ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE

Master's project in Environmental Engineering

Integration of Urban Structures in Point Process Analysis

Done by

Loïc Gasser

Under the supervision of

Prof. François Golay

In the laboratory of Geographical Information Systems
EPFL

External Expert Prof. Dr. Pierre Dessemontet

LAUSANNE, EPFL 2011

Table of Contents

1 Abstract

Traditionally, spatial analysis of point pattern has been mostly focused on Euclidean space. As many human related phenomena take place on a network, the assumption of a continuous isotropic space fails to describe events which actually occur on a one-dimensional subset of this space. Thus, recently, researchers have begun integrating network structure constraints to study point patterns. The focus of this report is primarily aimed at the integration of the network structure constraints in studying the first order property of point processes with Kernel Density Estimation (KDE).

Two different approaches and the computational methods used to calculate network based kernel density estimation (NetKDE) are described, and are then compared to each other as well as to KDE. An original approach which aim is to replace the conventional search area in flat disk through Euclidean space is introduced. In urban context, polygons of various shapes can be generated and used over the network as an approximation of the potential accessible area for a given distance.

As a first case-study, network based density values for various types of economic activities are generated for each building in Geneva. The integration of urban structures in the characterization of neighborhood attributes is an innovative approach which possesses many advantages. A classification based on the attributes generated with this method is performed, and a detailed analysis of the results is carried out.

In a second case-study in urban environment, time is considered as an additional dimension in kernel density estimates. A three dimensional KDE approach is used in an attempt to monitor the risk associated with the explosions of improvised explosive devices (IED) in Baghdad through space and time. An animation of the simulations is presented as a visualization technique to detect sensitive areas.

2 Introduction

As the availability of geographic data and the computers processing power is growing, new fields of investigations emerge to improve the process of knowledge acquisition. The integration of the city infrastructures in spatial analytical techniques is a very promising field of research. This enables to better explore, analyze and understand the underlying realties of urban context with large datasets. Nevertheless, solutions provided by commercial geographical information systems (GIS) software are often too limited in terms of flexibility and functions to address specific and complex issues. Moreover, the possibility for managing great amount of data in these environments is often a limiting factor. Therefore, there is a need to develop new flexible and user-friendly tools which can address these new challenges.

In urban context, thus far, road networks have been mainly used in transport planning. In this project, it will be shown how the integration of the network structure constraints in the analysis of point patterns can be used to characterize different phenomenon in the city. In urban context, the points can represent a wide variety of different events (economic activities, street intersections, burglaries, socio-economic surveys, traffic-accidents, trajectories…), and therefore their study can be used in numerous applications. Point pattern analysis represents one of the most interesting and adopted spatial analytical technique, because it is both easier to use and represent (Murgante, Borruso, & Lapucci, 2009). By taking into account the city's morphology, the model developed during this project focuses on the characterization of the physical environment at the scale of the neighborhood (up to 1 kilometer). As a result, the indices developed depend on the number of events located in a region (first order effect). Once the physical environment of an area has been determined with a number of pre-selected factors, it is possible to study the interactions between different events' attributes. This was performed by using a data mining method, but other approaches could be implemented. Ultimately, the goal of these analyses is to provide pragmatic spatial information to support the decision making in urban context.

The temporal dimension is often treated as an additional attribute of the events, but it will be demonstrated that time can also so be represented as a third coordinate in the determination of an event location. The spatio-temporal analysis of the first order properties of point processes is a new field of research, and thus far, has found applications in crime mapping and the study of trajectories.

3 Problematic and goals

A first study on NetKDE has already been carried out at LASIG, and the tool has been implemented notably to study the density of economic activities in Barcelona (Produit, Lachance-Bernard, Strano, Porta, & Joost, 2010) . Nevertheless, few insights have been provided to thoroughly explain the implications on the resulting density surfaces. Furthermore, the differences between NetKDE and a conventional KDE method have not concisely delineated thus far. Therefore, the first goal of this master is to

1) Explain the implications of using NetKDE over KDE

The original approach developed to calculate network kernel density estimations showed a high level of discontinuities in the resulting simulations. This is the result of the fact that previously developed technique implicitly assumed that Euclidean space was divided into fixed subsets according to the network configuration. As a result, the second goal of this work is to

2) Develop a new algorithm to increase the smoothing capabilities and reduce the artifacts created with this methodology

One of the major problems in the implementation of the original algorithm was that each computation was highly time consuming. In addition, all calculations needed to be repeated every time a parameter was changed. Thus, the third goal of this master thesis is to

3) Develop an algorithm which would avoid repeating the same calculations, and would therefore optimize the computation cost of repeated simulations on the same dataset

Once the new algorithm is implemented, the aim is to

4) Test the new approach in a real urban environment

Finally, the temporal dimension in KDE is a very interesting factor, which has been treated in only few scientific studies. Therefore, the final goal of this work is to

5) Present a methodology to integrate the temporal dimension in KDE and test this model to assess the associated danger of terrorist attacks in Baghdad over space and time.

4 Theoretical background

This chapter reviews the principal studies which are relevant to the present work, as well as the scientific background necessary to understand the subsequent analysis. Firstly, the kernel density estimator in Euclidean space is presented. Secondly, the main network features used to compute the different indices in a network space are introduced, together with insights of a few scientific studies. Thirdly, the different methods used to integrate the temporal dimension in KDE are reviewed. Finally, the clustering method used in the case-study of the city of Geneva is briefly described.

4.1 Kernel density estimation

This section focuses on the description of KDE. First, the general features of this method are introduced. Secondly, a description of the kernel function is presented. Thirdly, the importance of the bandwidth selection and the impact on the resulting density values are discussed. And finally, the biases introduced at the edges of the study area are summarized.

4.1.1 Generalities

Spatial analysis of point patterns can be classified in two categories:

1) Methods examining the first order effects of a spatial process, when it depends on the number of events located in one region.
2) Methods examining the second order effects of a spatial process, when it depends on the interactions among events (Murgante, Borruso, & Lapucci, 2009).

Methods which analyze the first order properties describe the way the mean value of a process varies across space (Xie & Yan, 2008). Amongst them, one can mention the quadrat analysis and Kernel Density Estimation (KDE). When utilizing the former approach, one loses information relating to the absolute location of the data points. Moreover, the resulting density values are very sensitive to the size, position and orientation of the chosen gird. KDE is a distance-based method which enables using the absolute position of events. Events are defined as spatial occurrences of the considered phenomenon, while points are each other arbitrary location (Murgante, Borruso, & Lapucci, 2009). Methods for analyzing the second order properties describe the way spatial events interact, and include methods such as nearest neighbor statistics, G function and K function (Xie & Yan, 2008).

Initially developed by Silverman in 1986, KDE is currently used in various fields including epidemiology, ecology, criminology and economics. A moving three dimensional function is used and centered on a number of locations in turn, usually a grid, to weight the points within its sphere of influence according to the distance at which the points are estimated (points that are further away are weighted less than those that are close). Thus, this method is used to obtain smooth estimates of univariate (or multivariate) probability densities from an observed sample of points (Silverman, 1986).

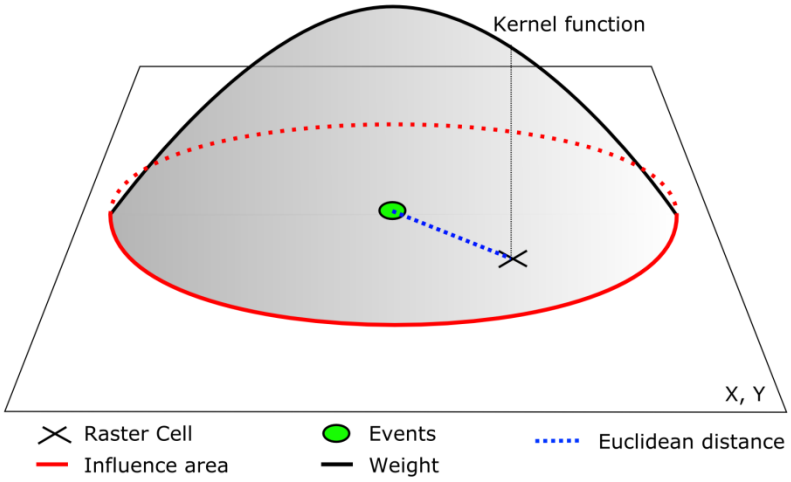

Figure 1: Influence area of an event with KDE (Produit, Lachance-Bernard, Strano, Porta, & Joost, 2010)

KDE can be applied to any dimension of Euclidean space, but when applied to geographical analysis, the points are usually in two-dimensional space. In order to perform kernel density estimation, one has to select a kernel function and a bandwidth (the maximal distance at which points are evaluated). Considering $\boldsymbol{x_j}$ a location vector over the field R and $x_1, \dots, x_n$ the locations of a set of n events, the intensity estimation is

$$\hat{f}_h(x_j) = \sum_{i=1}^{n} \frac{1}{nh^2} K\left(\frac{x_j - x_i}{h}\right) \qquad (1)$$

where $d_{i,j} = x_j - x_i$ is the distance between the gird point x_j and the event x_i and $K(.)$ is the kernel function (Silverman, 1986). The kernel density estimation can be thought of as a nonparametric, because rather than making assumptions about the distribution of a point pattern, it attempts to estimate it directly from the dataset (Brunsdon, 1995).

4.1.2 The kernel function

The kernel function is a technique similar to a classical weighting function which can take various forms, but must satisfy a number of properties. Each kernel function has the following properties:

1. $0 \leq K(s) < \infty$
2. $K(s) = K(-s)$
3. $\int_{-\infty}^{\infty} K(s)ds = 1$
4. $\int_{-\infty}^{\infty} K(s)\, s^2 ds = 1$
5. $\int_{-\infty}^{\infty} K(s)\, s^m ds < \infty$ for $0 \leq m < \infty$

The most commonly used kernel functions are the Gaussian and Epanechnikov functions. The one used in this paper is the multivariate Epanechnikov function, which is defined in d dimensions as

$$K(s) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2)(1-s^2) \; if \; s < 1 \\ 0 \; otherwise \end{cases} \quad (2)$$

where c_d is the volume of the unit d-dimensional sphere with $c_1 = 2$, $c_2 = \pi$ and $c_3 = 4\pi/3$.
The Epanechnikov function can be considered as optimal, because it minimizes the asymptotic integrated squared error (AMISE).
Therefore, in 2 dimensions, the formula becomes

$$K(s) = \begin{cases} \frac{2}{\pi} (1-s^2) \; if \; s < 1 \\ 0 \; otherwise \end{cases} \quad (3)$$

where $s = \frac{\sqrt{(x-x_i)^2+(y-y_i)^2}}{h}$. It should be noted that for the simulations presented in this work, the kernel function used is unimodal. The unimodal Epanechnikov kernel function in 2D is presented on Figure 2 hereunder.

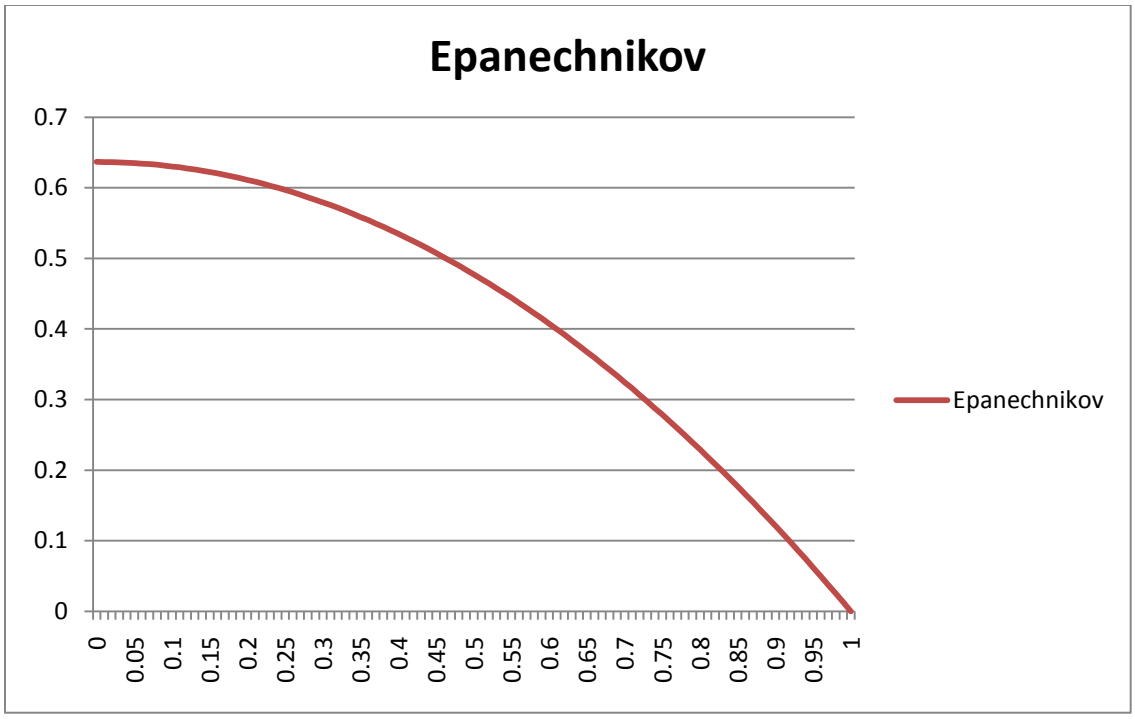

Figure 2: Epanechnikov unimodal kernel function in 2D

4.1.3 The bandwidth selection

The choice of the bandwidth, also called the window width or smoothing parameter by some authors, is the most determining parameter in the calculation of the kernel density values. The choice of the bandwidth directly controls the variance of the density function. High values will lead to large variance, whereas smaller values will have the opposite effect (Brunsdon, 1995). In other words, the bandwidth controls the degree of smoothness of spatial variation in the phenomenon under study. Narrow bandwidths produce spiky peaks, which are highlighted in the final distribution, while wider bandwidths produce smoother surfaces. Therefore, several clusters at different levels can be highlighted by selecting different bandwidths. The appropriate selection of a window width can be the result of a physical distance having a specific meaning for the phenomenon under study. It can also represent an optimal tradeoff between an excessive smoothing and spikiness. A number of methods have been developed to help researchers selecting an appropriate bandwidth, but this topic will not be discussed in the present paper. The bandwidth choice is the most important parameter for KDE. Its influence on the generated density surface is much greater than the choice of the kernel function (Xie & Yan, 2008).

In order to counteract the high variability of points' density within the same sample region, an author suggests using an adaptive kernel algorithm (Brunsdon, 1995). This method reduces the effects of over-smoothing areas where the point's density is lower and thus allows for a more representative and detailed image of the phenomenon being studied.

4.1.4 The edge effects

There are mainly two types of edge effects. The first one occurs when the subset of a wider area is extracted and only the events belonging to this division are selected. As a result, the events situated next to the study window are not taken into account and the resulting densities are therefore lower. The second potential edge effect occurs when the distance between an event and an edge is smaller than the bandwidth. In this case, the entire impact of the event through space is not taken into account, introducing inaccuracies regarding the relative density values of the other cells of the subset. Therefore, it is always better, if possible, to compute KDE in a window wider than the area of interest in order to avoid the edge effects.

4.2 Shortest-path tree

The shortest-path tree (SPT) represents the extraction of a part of a network. From a location vector located on a segment, using Dijkstra's algorithm (see Appendix section 11.1), all the possible shortest-paths are calculated until a limit distance (the bandwidth) is reached, then all the associated segments are selected to create the SPT. Figure 3 illustrates this concept, where part of the network is selected from a given centroid.

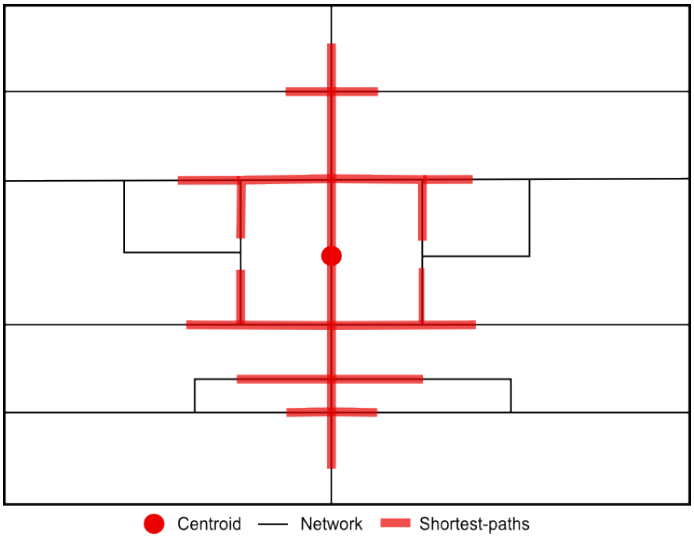

Figure 3: Shortest-path tree (SPT)

4.3 Accessible network-constrained surfaces

The concept of accessible surface calculated on a network is a very popular technique in Geographical Information Systems (GIS). Most studies use Euclidean distances to represent the potential accessible area from a starting point. Borruso (2008) uses a polygon to define a network-based area which is computed on the basis of the SPT. Thus, the polygons take various shapes and change according to the network structure. Figure 4 shows the difference in the search area for Euclidean and network distances.

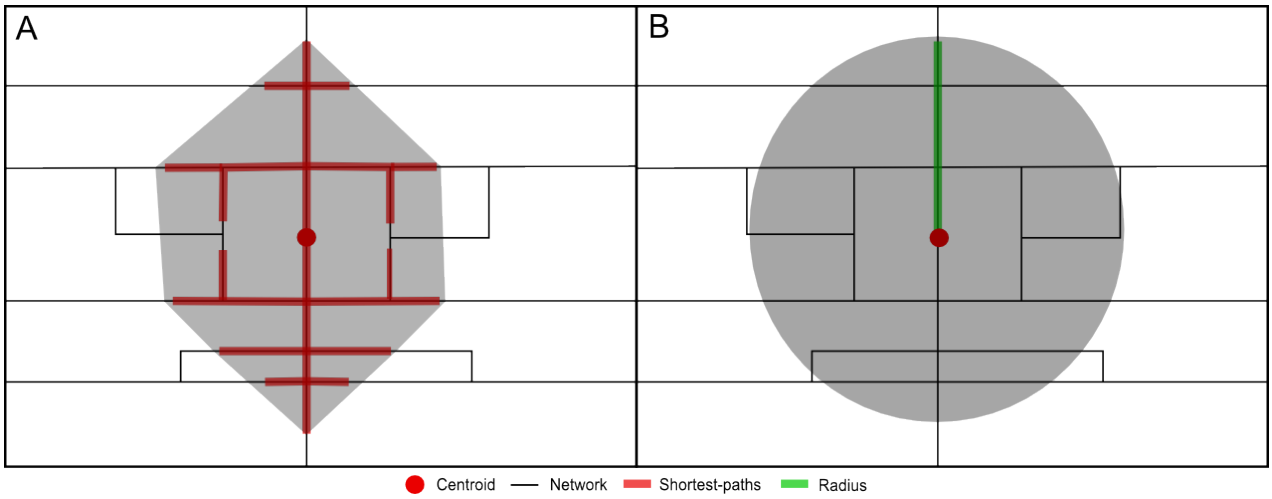

Figure 4: a) Search area for Euclidean distances (within a circle) b) Search area for distances on the network, convex hull type (within a polygon)

From all the ending points of the SPT on the network, the accessible area is computed by using a convex hull function. The convex hull of a set of points is defined as the minimum convex geometry that encloses all the points within the set, and the resulting shape is known as the minimum convex polygon (MCP). This concept was used in various studies to determine the market potential and influence of retail stores (Biba, Des Rosiers, Theriault, & Villeneuve, 2006),(Biba, Thériault, & Des Rosiers, 2007), (Biba, Thériault, Villeneuve, & Des Rosiers, 2008). Nevertheless, in the context of this study, distances are too high to assume that only the physical distance matters, therefore travel times are calculated for each trip using TransCAD (Geographic Information System for transportation), a software using street network and its constraints (orientation of the tracks, velocity limits, bends penalties …).

This technique has also been used in a slightly different way to determine the walking potential of few cities from each node of the street network (Genre-Grandpierre & Foltête, 2003). In this case, instead of using convex forms to measure the accessible area, a buffer of 5 meters along the shortest-path tree is generated. Using detailed studies on walking patterns within the cities of Lille and Besançon, it has been possible to identify the impact of the network morphology on mobility behaviors.

In this study, it will be shown how the concept of accessible areas can be extended and used for various purposes. Indeed, this simple approach seems to better capture the underlying realities of displacements within a network constrained city.

4.4 Network density

Borruso (2003) developed a network density estimation based on the number of junctions falling within a predefined grid. The results are first derived as an absolute number of junctions per cell, followed by a more refined analysis based on the Kernel Density Estimation of the first results. KDE allows for a better visualization of the densities and is less dependent on the size, position and orientation of the chosen grid than absolute densities. It has been demonstrated that for the city of Trieste, clusters of built areas usually match with clusters of network densities. Borruso (2003) also highlighted the fact that by selecting different subsets of the road network, it was possible to distinguish different cores in the city based on the user's perspective.

In another article, Borruso (2008) introduced the concept of Network Density Estimation (NDE). The aim of this method is to calculate densities of point patterns by taking into account the fact that space within a city is constrained by the network. First, a fine resolution grid is generated above the study area. A buffer is then defined around the network, and only the cells of the grid belonging to this interval are kept to estimate the densities. Areas defined by the shortest-paths tree are then used to calculate the absolute density of events intersecting them. NDE densities can then be expressed in terms of both 'events per kilometer' (the overall length of the segments) and 'events per square kilometer'. As the densities are only computed on the cells' centroids close to the network, values are interpolated using an Inverse Distance Weighting method to obtain a continuous surface. This method has been tested on densities of banks and insurance companies in the Central Business District of Swindon. It has been demonstrated that NDE seemed more effective at highlighting linear clusters oriented along the street than KDE.

A method which only calculates densities on the network has been developed to estimate the density of traffic accidents (Xie & Yan, 2008). The key feature of this approach is that the network space is represented with basic linear units called lixels. Therefore, densities are calculated per linear unit instead of per area unit. Thus, densities are only estimated on the network and not on the region where the network is embedded. It has been emphasized that using a conventional KDE method tends to cover space beyond the network context and over-estimate the density values.

Other authors pointed out the fact that when dealing with density of points on a network, a bias due to the type of kernel function applied is introduced. Indeed, they demonstrated that when the degree of a node is greater than or equal to three, space in not homogenous and therefore, new forms of kernel functions on networks should be used. Based on a number of properties a kernel function on the network should satisfy, three types of kernel functions have been formulated (Okabe, Satoh, & Sugihara, 2009). However, none of the formulated modified kernel functions satisfy all the properties, but two of them can be considered as unbiased.

4.5 Network kernel density estimation

In this section are presented the methods which attempt at calculating densities per area unit. The challenge is to find an appropriate method which enable to compute densities for the entire region where the network is embedded. Therefore, distances used in the kernel function are always a combination of Euclidean distances and distances measured on the network.

4.5.1 The closest node approach

NKDE (network kernel density estimation) is a methodology in which each grid point is connected to its closest node (Downs & Horner, Characterising Linear Point Patterns, 2007). The shortest-paths to the events are then computed from these nodes and a kernel function is applied to calculate the resulting densities. Thus, the main difference between this approach and the ones introduced earlier is that values are no longer expressed as absolute densities. This method has been applied to car accidents and animals home ranges (Downs & Horner, 2007). Nevertheless, the datasets used in these calculations are much smaller than the ones used for the simulations presented in this work. One of the major drawbacks of such an implementation is that, in a loosely connected network with long segments, the selection of the nearest node can introduce important biases in the calculation of the densities. Furthermore, high densities will tend to cluster around the network's nodes.

4.5.2 The closest segment approach ("NetKDE closest")

As most human related phenomenon usually take place in a space constrained by a network, Euclidean distances measured in an isotropic space often fail to represent the true distances travelled along the network. According to the network morphology, important sources of inaccuracies can lead to misinterpretations of the results.

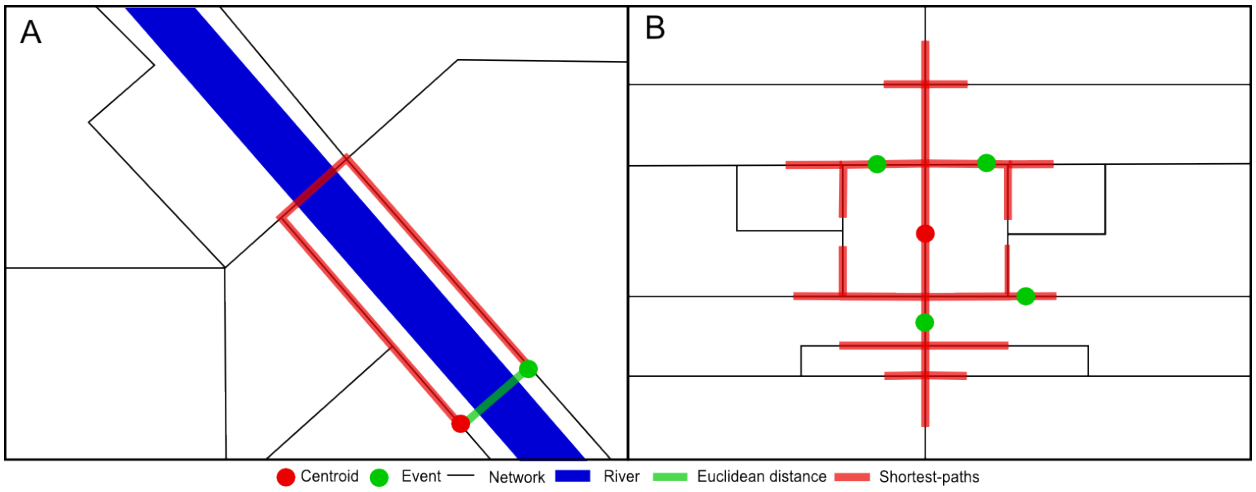

Figure 5: a) Effect of a bridge on the shortest path b) Events selection with the SPT for NetKDE

Figure 5 a) is a good example of the potential differences between the Euclidean distance and the shortest-path. One is forced to take the bridge in the North to be able to reach its destination. Produit *et al.* (2010) introduce a kernel function applied to distances measured on the network. Distances are weighted the same way as a conventional KDE, but instead of looking for the events falling within the radius of the circle defined by the bandwidth, they are selected according to their belonging to the SPT. Figure 5 b) shows a representation of the event selection with the SPT. The network kernel density estimation (NetKDE) equation is

$$NetKDE(x_j) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{d_{net,ij}}{h}\right) \quad (4)$$

where n is the number of events belonging to the SPT, x_j the centroid we are looking at, h the bandwidth, $d_{net,ij}$ is the distance between the centroid j and the event i, and $K(.)$ represents the kernel function which , as previously explained, can be of various types.
In order to compute NetKDE, a grid of points is generated above the study area, and each centroid is projected on the closest segment. In order to take into account the extra effort needed to reach the segment and to smooth the resulting map of densities, the distance between the segment and the centroid is also taken into account. This method has been already implemented to study the density of economic activities in Barcelona (Produit, Lachance-Bernard, Strano, Porta, & Joost, 2010). This study points out the fact that using NetKDE reflects more adequately the reality of urban mobility. Another study used this approach to determine the optimal location for cycle paths and lanes development (Lachance-Bernard, Produit, Tominc, Matej, & Golicnik Marusic, 2011). The data used to perform NetKDE were point events collected by GPS devices and/or web-based GIS portals. The added value of using this method relies on the fact that the data are presented in a more understandable way, making it possible to transmit more easily the collected information to the planners.

4.5.3 The visible segments approach ("NetKDE visible")

As explained in the previous section, network kernel density estimates thus far have been calculated by selecting the nearest segment from the centroid being studied to create the SPT. This approach minimizes the distance in Euclidean space and, therefore, remains very strongly related to the network. The space is virtually divided into polygons according to their distance to the closest segment. The previous tests showed that segmentation of space in this manner led to highly discontinuous density values, which can be hard to interpret. As a result, a new approach called "the visible segments" approach has been implemented. In this method, the sum of the shortest-path and the Euclidean distance is minimized. In order to be able to calculate such a density value, the centroid is projected on all the surrounding ("visible") segments and all the shortest-paths are computed to the target event. On Figure 6, one can see the different paths produced by each approach. NetKDE visible allows for the projection of the centroid on various surrounding segments according to the location of the event. Therefore, Euclidean space is no longer artificially partitioned and therefore NetKDE is less dependent on the location of the centroid. Density surfaces are expected to be smoother than with the previous implementation. To our knowledge, this is the first time such a method has been used to calculate network kernel density estimates.

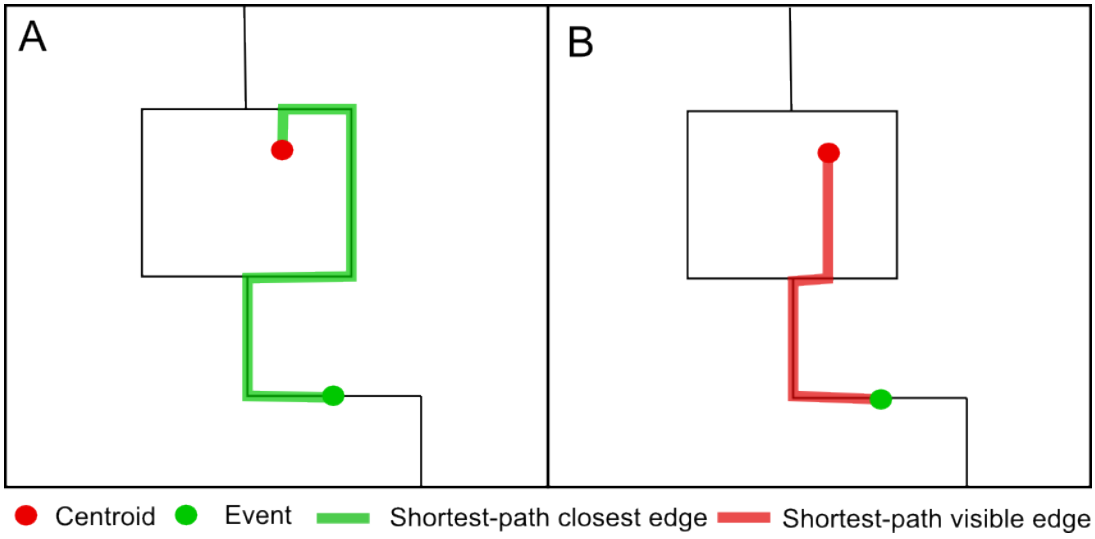

Figure 6: Comparison between the shortest-path computed with the closest segment approach and the shortest-path computed with the visible segments approach. a) Shortest-path computed with the closest segment approach b) Shortest-path computed with the visible segments approach

4.6 KDE and spatio-temporal analysis

In this section, the principal ways of integrating the temporal dimension to the KDE calculations are reviewed.

4.6.1 The dual KDE approach

The KDE can be applied to two variables, resulting in a dual density estimate and called dual KDE. The variables are first determined individually and then related to one another with simple algebraic operations. The dual kernel density estimation has been used, for instance, to study the spatio-temporal variations in concentrations of retail stores in the province of Upper Austria between 1998 and 2001(Jansenberger & Staufer-Steinnocher, 2004). In this study, an arithmetic difference has been performed using the absolute density values for both periods.
Other authors use the KDE to analyze burglary crime scenes (Wolff & Asche, 2010). In order to analyze spatio-temporal patterns, a surface for each month is generated in order to study monthly variability. Therefore, time is not directly used in the kernel function, but defined time intervals are set in which the associated crimes are selected.

4.6.2 The comap

The comap, which is derived from the term Conditional MAP, is a tool which can combine the spatio-temporal aspects of point patterns with kernel density estimations (Brunsdon, 2001). This exploratory data analysis tool is essentially a geographical variant of the coplot. As such, instead of using a plot, Brunsdon suggests using maps for visualizing data.

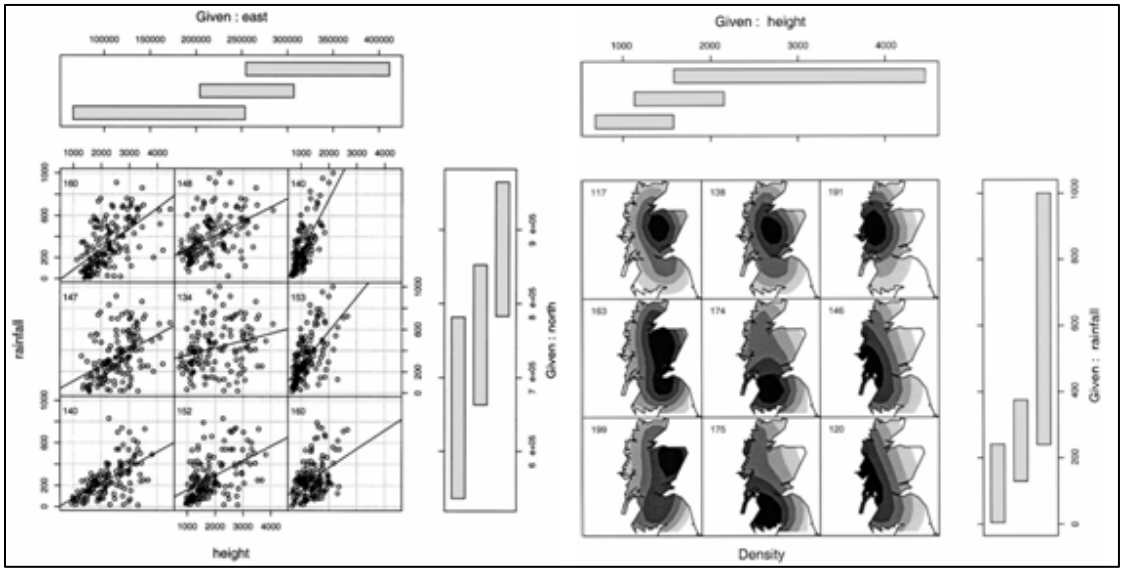

Figure 7: Coplot and Comap using two conditional variables (Brunsdon, 2001)

Figure 7 shows the traditional approach of using a coplot in order to visualize rainfall events. The two extra plots situated on the top and left of the plots show the ranges of each conditional variable, and, therefore, how the points are clustered in the six different plots. An important issue to address when generating a coplot, is the fact that it is very important to try to have the same number of observations on each scatterplot. This is why the ranges do not have the same size and are overlapping. Nevertheless, when dealing with more than one conditional variable, it is no longer possible to have the exact same number of events in each scatterplot. This issue is addressed here by the author by printing the number of observations in the top left hand corner.

Figure 7 shows how the same approach as the one used for the coplot, but now instead of using scatterplots, maps of event densities are used. The conditional variables here are height and rainfall, but it would be possible to use virtually any kind of variable. Some authors suggest using the comap to combine the spatial and temporal dimensions to study fire incidents in urban areas (Chhetri, Corcoran, & Stimson, 2009).

4.6.3 3D spatio-temporal KDE

4.6.3.1 Integrated spatio-temporal function

Demsar and Virrantaus (2010) generalized the concept of the standardizing a 2D KDE into a 3D spatio-temporal KDE. To do so, they used a common cartographic visualization method, which is known as the space-time cube and was first introduced by Hägerstrand in the 1960s. The x and y axis represent the geographical space (e.g. a plan), and the z axis the time. This approach has been used to study the trajectories of vessels from the Gulf of Finland, represented as points, and connected to each other according to their occurrence on a time line. In order to calculate 3D dimensional densities, a space-time cube regularly divided through space and time into voxels (e.g. equivalent to a pixel in 2D) has been used. The densities of each trajectory have been calculated for one trajectory at a time and then summed up to obtain the total density for the entire dataset of trajectories. The value of each voxel is computed according to the distance from its central point to the trajectory. Therefore, the same kernel function is applied to space and time. Implicitly the kernel function can be expressed as

$$\hat{f}(x,y,t) = \frac{1}{nh^3}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}, \frac{y-y_i}{h}, \frac{t-t_i}{h}\right) \quad (5)$$

where h is the bandwidth in 3D and n is the total number of events. This means that the same kernel function is applied to space and time and that all the parameters interact with each others. From a given point (e.g. event) in 3D, densities are distributed over a sphere (Figure 8).
The authors suggest different visualization methods, which include the direct volume rendering, isosurfaces and volume slicing or clipping with planes. It has been demonstrated that this technique is very efficient to explore and analyze large datasets of moving objects. Nevertheless, it does not reveal information on other attribute movements such as speed and direction.
The Epanechnikov kernel function in 3 dimensions (Silverman, 1986) is expressed as

$$K(s) = \begin{cases} \frac{15}{8\pi}(1-s^2)\ if\ s < 1 \\ 0\ otherwise \end{cases} \quad (6)$$

where $s = \frac{\sqrt{(x-x_i)^2+(y-y_i)^2+(t-t_i)^2}}{h}$. Therefore, the distance from an event to a point within its sphere of influence is nothing but the norm of the vector.

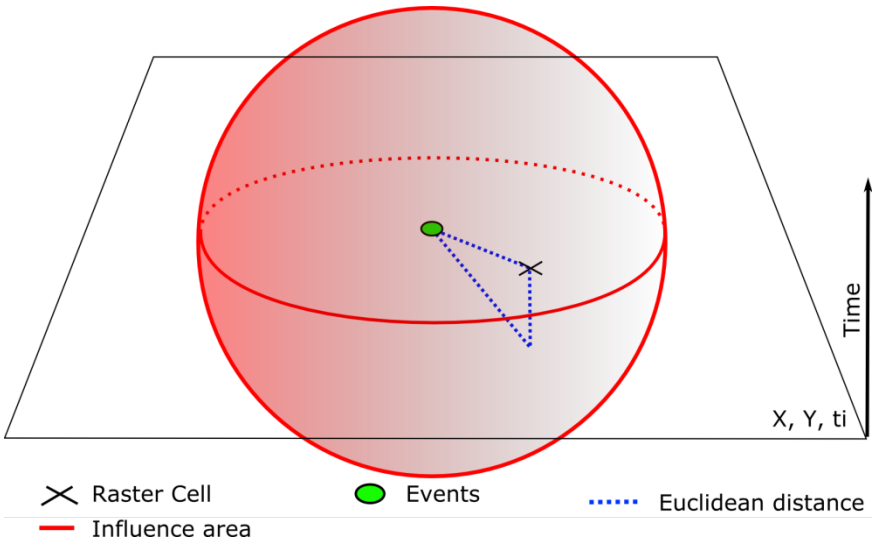

Figure 8: Volume of influence of an event using a 3D integrated KDE approach. Time is considered on the z axis.

4.6.3.2 Separated spatio-temporal functions

Other authors suggest using a different function for the integration of the temporal dimension in kernel density estimates (Brunsdon, Corcoran, & Higgs, 2007)(Nakaya & Yano, 2010). The formula can be expressed as

$$\hat{f}(x,y,t) = \frac{1}{n{h_1}^2 h_2} \sum_{i=1}^{n} k_1\left(\frac{x-x_i}{h_1}, \frac{y-y_i}{h_2}\right) k_2\left(\frac{t-t_i}{h_2}\right) \quad (7)$$

where $k_1()$ represents the kernel density over space with the bandwidth h_1, and $k_2()$ represents the kernel density over time with the bandwidth h_2. In this case, for a given point in 3D, densities are distributed over a cylinder (Figure 9). This function enables to set fixed bandwidth values over space and time. Nevertheless, the resulting density values are generated by the product of two kernel functions, and, therefore this method is a hybrid version combining 2D and 1D kernel functions. This approach has been used to visualize crime clusters in a space-time cube (Nakaya & Yano, 2010). It has been highlighted that this method enables an effective simultaneous visualization of the geographical duration of crime clusters. Compared to a traditional crime mapping using arbitrary fixed time intervals, this approach facilitates the visualization of geographical diffusion and movements of crime clusters.

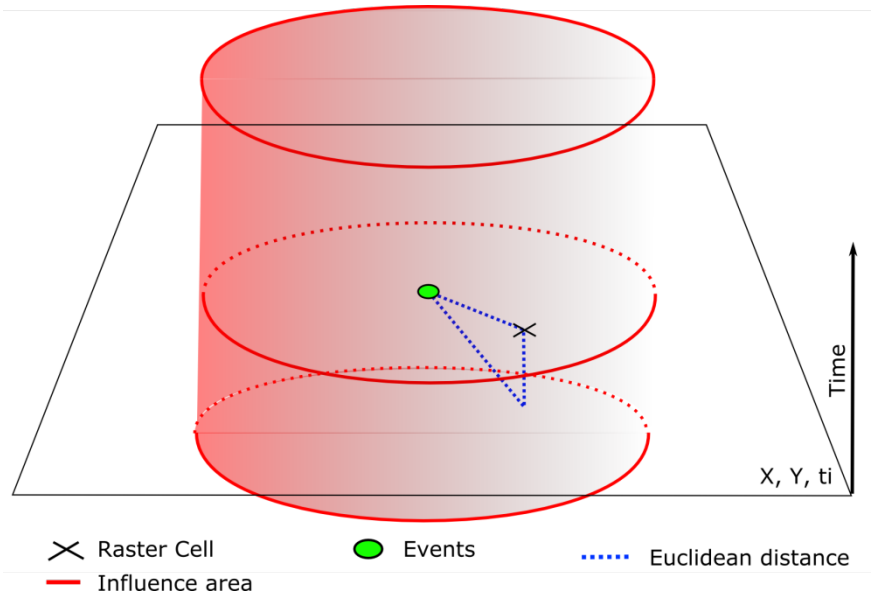

Figure 9: Volume of influence of an event using a 3D separated KDE approach. Time is considered on the z axis.

4.6.3.3 Visualization approaches

4.6.3.3.1 Isosurfaces

Other authors have proposed the use of isosurfaces as an effective visualization technique(Brunsdon, Corcoran, & Higgs, 2007)(Demsar & Virrantaus, 2010). These surfaces are calculated from the 3D scalar field similarly to 2D surfaces, which share the same scalar field values. Isosurfaces provide a way to visualize the whole time period under study. Nevertheless, problems arise when trying to show more than one value of density. Indeed, often isosurfaces enclose each other and hide parts of the information.

4.6.3.3.2 Volume rendering

The volume rendering method is considered as an extension of the isosurface approach, but while the latter enables the visualization of only one density at a time, the volume rendering method enables the simultaneous visualization of different densities by controlling opacity and light effects (Nakaya & Yano, 2010).

4.6.3.3.3 Volume slicing

Sections of the volume can be cut with planes, usually horizontally or vertically, and allow for the visualization changes in densities through space or time. This technique has been used to explore and visualize ground-motion data from an earthquake (Hsieh, Chen, & Ma, 2010). Others have used it to study trajectories of vessels in the Gulf of Finland (Demsar & Virrantaus, 2010). This approach is very interesting because it allows the user to choose the visualization with greater flexibility.

4.6.3.3.4 Animation

The animation visualization approach is nothing more than a regular volume slicing through time. To my knowledge, there have never been any publications on animations using 3D spatio-temporal KDE. This approach is intuitive and very useful to identify key epochs (Brunsdon, Corcoran, & Higgs, 2007). Nevertheless, in order to be really efficient, the user should be able to interact during the continuous sequence of snapshots. Indeed, the potential problem of using animations an exploratory tool comes from the user ability to remember all the pertinent information of the animated sequences (Brunsdon, Corcoran, & Higgs, 2007).

4.7 Diversity

A first, very simple measure of diversity is the **richness** m which represents the total number of different types of spatial actors in the study area. Nevertheless, this indicator does not provide much information about the distribution within a sample. A better indicator, known as the Shannon diversity index and used initially in ecology to measure biodiversity, provides information about the overall **entropy** of a system. With N_i being the number of actors belonging to the class i and N the total number of actors, the diversity index is calculated as

$$H = -\sum_{i=1}^{N} p_i ln(p_i) \quad (8)$$

where $p_i = \frac{N_i}{N}$. Therefore, the entropy ranges from 0 (with only one type of activity in the sample) to $ln(m)$ (with an equal share of all the activities) (Van Eck & Koomen, 2007). One can also assess the level of **dominance** of a species over the other species of the system. The equation for this index is

$$p_{max} = max(p_i) \quad (9)$$

The level of dominance can be combined its associated class to determine which class is dominating where, and at which level.

4.8 Simple K means clustering method

K-means clustering is a very popular technique used in numerous applications. This method applies to a number of n observations divided in k classes. The observations possess d quantitative attributes in d-dimensional space. Given an integer k, the goal is to create k classes so as to minimize the mean squared of the distance from each observation to its center (also called centroid) (Kanungo, M. Mount, Silverman, Netanyahu, Wu, & Piatko, 2000). Initially, k observations are randomly selected and k classes are generated according to the Voronoi diagrams method. Then, the mean value of the each class is calculated and the virtual points at the center of each class are created. The classes are then updated according to their new centroid (attribute the observations to the k classes according to their nearest center). This procedure is repeated iteratively until some convergence conditions have been met. This method will be used later to determine categories of types of economical activities for the city of Geneva. The software used to perform this simulation is *CommonGIS*.

5 The visible approach implementation

The aim of this chapter is first to briefly introduce the structure of the original algorithm ("NetKDE closest") developed at LASIG and point out its weaknesses. Then a more detailed description of the new version ("NetKDE visible") will be provided, as well as a summary of the differences between the two techniques. Finally, the constraints as well as the choices made for the development of this tool will be explained.

5.1 Data preparation

The network must be prepared before the simulation can be started. This step is performed with a python script which accesses the ArcGIS toolbox using the module "arcgisscripting". The network's topology and geometry are repaired and cleaned. Nodes are created at each intersection and each time a segment ends without any further connection. Each node possesses a unique Id and is used in several sub processes presented in the following algorithms.

5.2 Original Algorithm

A first version of this tool has already been developed during the Master Thesis of Timothée Produit in 2010. It is important to point out that the version described later is slightly different from the one initially developed. The goal of the new implementation was to create a program which would exhibit an increased flexibility and performance for repeated simulations, and which would enable to calculate NetKDE densities using the visible segments approach. Nevertheless, in order to understand the advantages of the new tool, it is necessary to introduce the previous version to point out where the algorithm could be improved and what are the major differences between the two implementations.

5.2.1 Inputs

The inputs are the same for both versions of the tool. Before explaining the algorithm in detail, it is important to clearly define the terms which will be used during the description of the computation steps.

The word **centroid** is used to define the center of gravity of an object. For a raster cell, the centroid is located at the intersection of its diagonals. For the following sections, we will always be referring to a regular grid, but as it will be shown later, centroids can also be generated for different objects.

The **nodes** represent the intersections of the network's segments or the end of a segment. They all possess a unique Id and are needed to compute the shortest-path algorithm.

Finally, the **events** (activities or spatial actors) represent the point patterns under study, e.g. the phenomenon one is trying to gather information about.

Data Layer	Minimal attributes
Network	- One Id per edge - From Node Id - To Node Id - Geometry (coordinates, length)
Nodes	- A node Id per node - Geometry (coordinates)
Centroids	- An Id per centroid - Geometry (coordinates)
Events	- An Id per event - Geometry (coordinates) - Attributes (optional) - Weights (optional)

Table 1: Input layers

5.2.2 Computation steps

5.2.2.1 Creation of the grid points

First, it is necessary to create a grid of points for which the spatial resolution can be chosen freely. The user can either generate a grid using the extent of the network to which the length of the bandwidth is added, or define a point on the lower left corner of the area of interest and define the number of points that need to be generated for the height and width.

5.2.2.2 Extraction of a part of the network (from the centroids)

In order to avoid calculating the shortest-paths on the totality of the network, the segments which are intersecting (touching or within) the buffer generated around the centroid are extracted. The radius of the buffer is equal to the bandwidth.

On Figure 10, it is interesting to point out that segments going further than the bandwidth are selected to make sure the shortest-paths can be calculated on the network before reaching the limit set by the bandwidth.

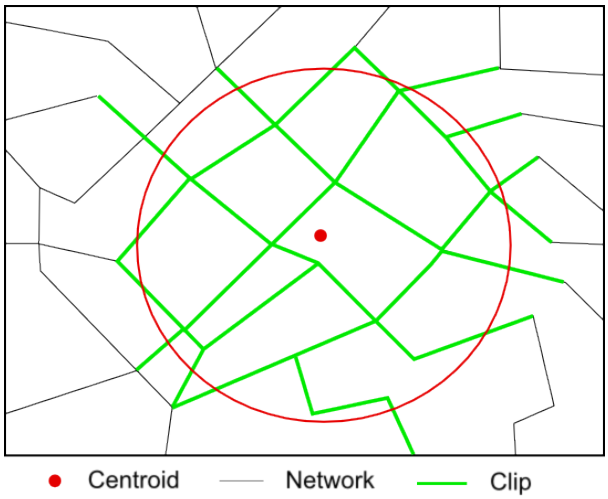

Figure 10: Clip of the edges (red circle)

5.2.2.3 Centroid's projection on the closest segment

Once the segments inside the buffer have been selected, one looks for the closest segment on which the centroid is projected. During this step, the closest segment is virtually divided into two smaller segments, which will enable the calculation of the shortest-paths in both directions. The distance between the centroid and its projection, as well as the coordinates of the projected point and the Id of the closest segment, are kept in memory.

As the creation of a buffer and the analysis of its intersection with the edges are computationally very demanding, the projections of the centroids on their closest edges are calculated before starting the simulations and stored in PostgreSQL.

5.2.2.4 Shortest-path algorithm

During this step, the Dijkstra's algorithm introduced previously is used to calculate the shortest-path tree (SPT). Therefore, all the segments belonging to the bandwidth are selected.

5.2.2.5 Extraction of the events and selection of the closest segment

Once the shortest-path tree is computed, it is possible to select the events projected on the SPT for a given bandwidth. As it has been done for the centroids, the events' projections on their closest segments are calculated as an upstream process, where their projections' coordinates and closest segments Id are kept in memory.

5.2.2.6 Computation of the kernel density

Once the segment on which the events are projected is known, it is possible to compute their distances to the centroid. The weights of each event are then calculated using a kernel function, which has previously been introduced. The sum of all these weights divided by the bandwidth squared produces the final network density estimation for a given centroid.

5.2.2.7 Limitations of the first version

One of the limitations of this algorithm is that it requires the calculation of the SPT from each centroid. Therefore, in order to obtain a high resolution, many similar shortest-path trees must be computed. Moreover, if a parameter is changed, all the calculations need to be repeated after each simulation.

For the same network, a list of the parameters one might want to change is provided below:

- The bandwidth
- The kernel function
- The selection of a specific kind of event
- The selection of different time period
- The use of different centroid layers (not only a grid point)
- The introduction of new weighting factors (as an attribute of the event for instance)

Another limitation relies on the fact that it is impossible to specify whether the distances centroid-segment and/or event-segment must be taken into account or not. Furthermore, the visible segment approach introduced earlier has not been implemented in the first version of the algorithm.

5.3 New algorithm (visible segment approach)

In this section the main features of the new algorithm will be described in more details. Instead of starting the calculations from the centroid, the shortest paths are first defined from one node to every other node of the network for a given bandwidth. Moreover, the segment selected to calculate the shortest-path to the event is not necessarily the closest one.

5.3.1 Choice of the platform

The platform's choice can be explained for several reasons. First, the implementation of the new algorithm required a database in order to store intermediate calculations. PostgreSQL combined with the spatial extension PostGIS is a powerful and open source database system. It has more than 15 years of development and has been shown to be very reliable. Moreover, it can be integrated in a lot of programming languages (including python) and posses an exceptional documentation. The first algorithm was initially developed in python, which is also an open source programming language. Python is easy to learn and use, which is a very important feature when a user has little or no programming experience. More modules are constantly becoming available and the python community is growing every day. Thus, it was decided to keep the same platform, which has already showed good results during the development of the first algorithm.

5.3.2 User's interface and parameters

The interface has been built with the module *wxPython*. On Figure 3 below, the actual interface is presented.

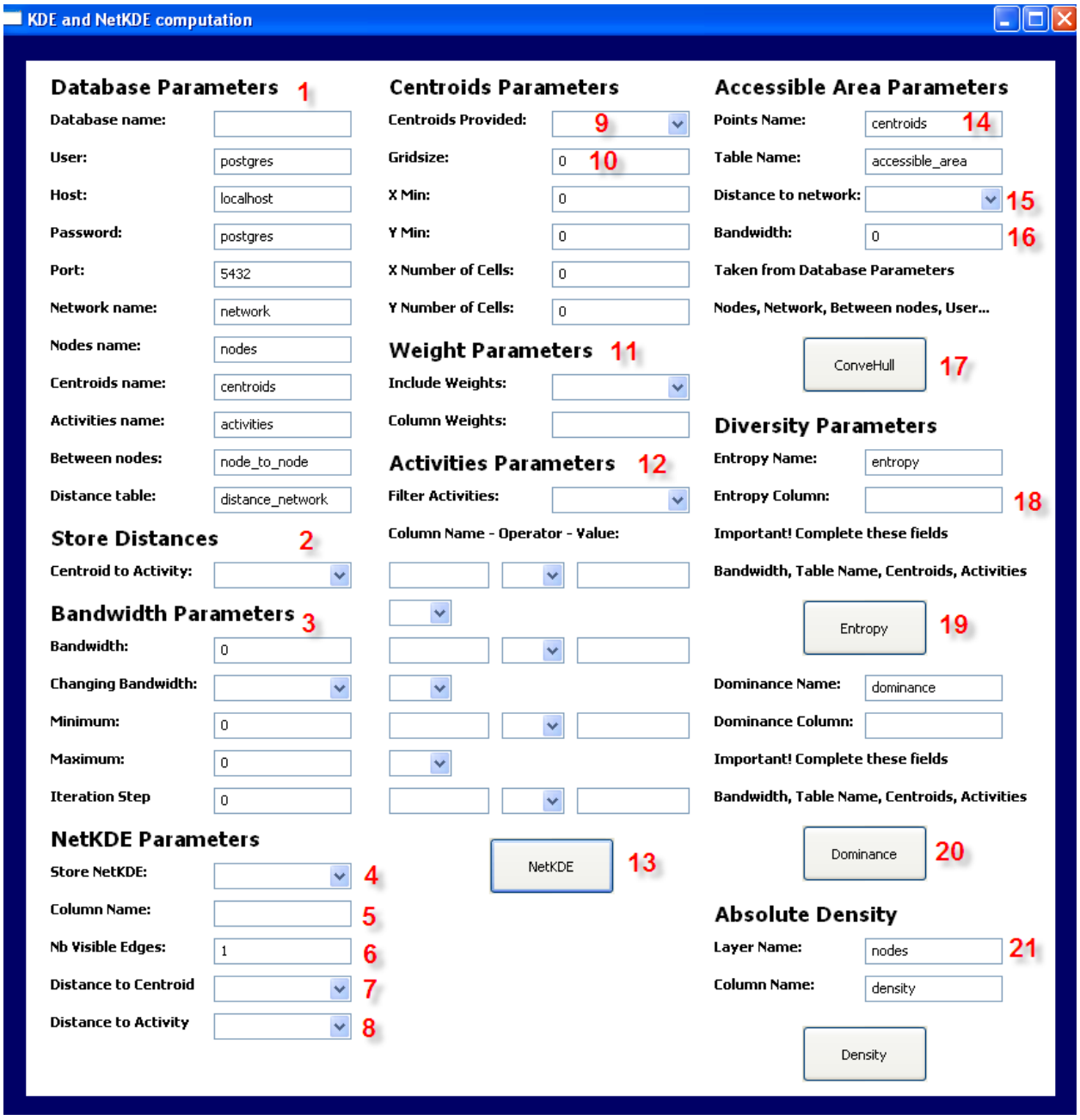

Figure 11: Interface of the program

1. Specify the database parameters.
2. Define whether or not to store the distances between the centroids and the activities.
3. It is possible to ask the program to calculate NetKDE values for several different bandwidths at the same time. The algorithm will always start with the highest bandwidth and then check iteratively the values for the smaller ones.
4. Define whether or not to compute and store the NetKDE values.
5. Define the name the NetKDE column(s).
6. Define the number of segments you want to analyze in computing NetKDE values.
7. Define whether or not to take the distance between the centroid and its projection into account.
8. Define whether or not to take the distance between the event and its projection on the network into account.
9. Define if a centroid layer will be provided.
10. If the following parameters are left as they are, the program will automatically create a grid of points; otherwise the extent can be set manually.
11. Define whether the spatial actors must be weighted, and if it is the case, what is the name of the column to be considered.
12. The events can be filtered according to certain kind of attributes. Define first whether or not to filter the events and then define the specific attribute's names and values, as well as the connectors between them.
13. Compute NetKDE values according to the previous parameters
14. Define the layer on which the accessible areas must be generated. If the distances between the nodes are not stored in the database, there are computed before to generate the polygons.
15. Define whether to take the distances between the points and their projections into account.
16. Define the bandwidth for the computation of the accessible surface.
17. Run the simulation. It must be noted that the database parameters are taken from the previous section. This includes all the connection parameters, and the layer name of the network, the nodes and the table "node_to_node".
18. Select the column from the table "activities" on which the entropy must be measured.
19. Compute the entropy. The name of the polygon layer must be introduced above as well as the corresponding bandwidth. The entropy value is generated per default on the centroid layer.
20. Compute the dominance index.
21. It is possible to measure absolute densities within the polygons and any point layer can be used.

5.3.3 Computation steps for NetKDE

5.3.3.1 Creation of a grid of points or provide a centroid layer

A grid of points can be generated in the same way as described for the first algorithm or a layer containing any kind of centroids can be provided. A GiST index is created on the geometry column so that subsequent spatial queries are accelerated.

5.3.3.2 Extraction of a part of the network (from the nodes)

The main difference between the original and the new algorithm is that instead of extracting part of the network from the centroids, the segments are extracted from the nodes. As previously explained, the creation of a buffer and the analysis of its intersection with the network's segments appeared to be computationally very demanding. Therefore, it has been decided to use a square rather than a circle. This resulted in a significant decrease in computing time, which is due to the fact that a square is a less complex shape and that it enables the use of the GiST (Generalized index Search Tree) indexes of the segments' geometries. The sizes of the box's segments are equal to the bandwidth and the nodes are successively at its center.

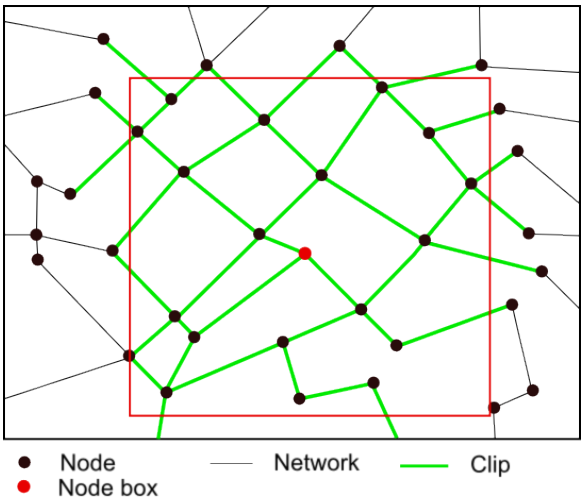

Figure 12: Clip of the edges (box)

5.3.3.3 Shortest-path algorithm

Dijkstra's algorithm is then used to determine the shortest-paths from one node to all the nodes associated to the selected segments. During this step, it is possible to create a table which contains three columns; there are two columns to store the combination of node's Id and another to store the distances between the nodes. A dictionary keeps track of the nodes visited in order to store the distances between the nodes in only one direction. Once this table has been generated, a b-tree index is created at the same time on both columns storing the node's Id, because these two attributes are always accessed jointly. The creation of indexes becomes an absolute necessity when dealing with large databases.

5.3.3.4 Event's selection

Before projecting the events on their closest visible segment(s), it is computationally advantageous to make sure events are in the vicinity of the centroid considered. Once again, a bounding box is used to enhance the retrieving velocity.

5.3.3.5 Centroid's projection on the closest visible segment(s)

Once all the distances between the nodes of the network have been computed in a given bandwidth, the centroids are projected one after the other either on their closest segment, or on their closest visible segments using a box to reduce the computation cost.

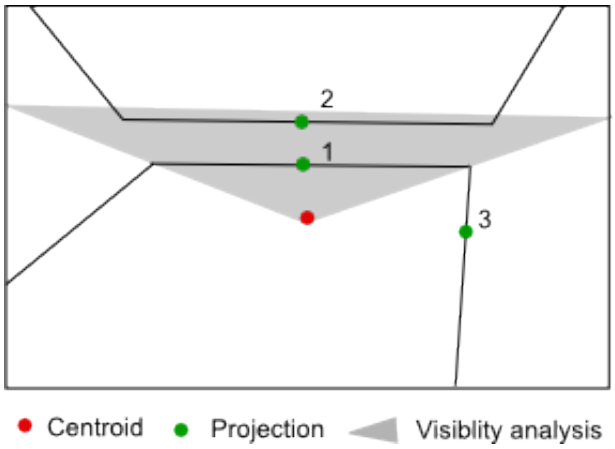

Figure 13: Visibility analysis

If the centroid can be projected on several segments (user's defined parameter), the program must make sure that the segment on which the centroid is projected is visible from its location and is not hidden behind other lines. Thus, once the segments are selected, they are classified in ascending order according to their distance to the centroid. Then, the closest edge is taken first and the interval angle constituted by both segment's vertexes (according to the azimuth) is kept into memory. The next segment is then analyzed and it is checked to be sure it falls within the interval defined by the first line. If the segment is not within this angle, it is kept into memory and its visibility is computed (angle), otherwise the next segment is analyzed.

In Figure 13, point 2 falls within the angle defined by the closest segment and therefore will not be taken into account, while point 3 is not within this angle and will be considered as a possible projection edge. A detailed flow chart of the visibility analysis is presented in Appendix 11.3.
During this step, the distances of centroid-segment(s), FromNode-projection(s) and ToNode-projection(s) are kept in a dictionary in python's memory.

5.3.3.6 Events' projection on the closest segment

Subsequently, the events which are within the box are selected and projected on their closest segment. The selection of the events can also be based on attributes which can be set before running the simulation. If several bandwidths are selected, they are sorted in decreasing order. The events' selection is performed only once with the highest bandwidth. Then, as soon as the distance to an event is lower than the bandwidth, it is removed from the events' list.

5.3.3.7 Shortest-path between a centroid and the selected events

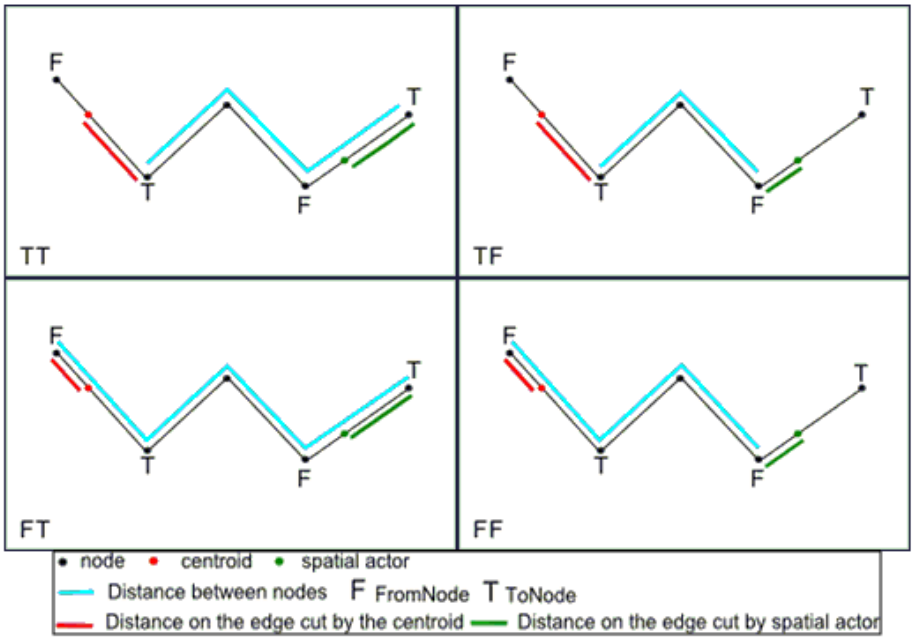

Figure 14: Computation of the Distance centroid-event

In order to compute the distance between a centroid and an event, one has to compare four distances which are illustrated on Figure 14. All the segments have a FromNode and a ToNode, which are created during the digitalizing process. Thus, they all possess an orientation and it is therefore possible to determine which node is associated with which part of the segment cut by the projections of the centroid and the event. Once the distances are computed, the shortest distance corresponds to the only real path between the centroid and the event. The only exception occurs when the centroid and the event are projected on the same edge. In this situation, the distance between the two points is computed by subtracting the maximum distance between the two projected points towards the same node with the minimum distance. If several segments were selected during the previous step, the distances on the network are computed from each segment and then compared to each other and the shortest one is kept. The advantage of storing the distances between the nodes is now more obvious. If one wanted to compare the distances between each centroid's projection and a selected event, one would have to go through the Dijkstra's algorithm for each projection.

After the distance between a centroid and an event has been determined, it is possible to store it in a new table in the database. This table possesses three columns, one for the centroid's Id, another for the event's Id and a last one for the distance. This functionality is particularly interesting to speed up the subsequent simulations. A b-tree index is created on the centroid's Id column which will speed up access to the database. A foreign key is also created on the event's Id column which is referenced to the initial event's table. A detailed flow chart of the distance computation is presented in Appendix 11.4

5.3.3.8 Computation of the kernel density

The computation of the kernel density estimation is processed for one centroid at a time and for all the bandwidths. A column in the spatial actor's table can be selected in order to assign an extra weighting factor. With $d_{net,ij}$ being the distance between the centroid and the event, h the bandwidth and w_i the weight of the event i; the NetKDE intensity at the centroid's location vector x_j over the field R is

$$NetKDE(x_j) = \sum_{i=1}^{N} \frac{1}{nh^2} K\left(\frac{d_{net,ij}}{h}\right) w_i \quad (10)$$

Thus, weights can be used in a situation where events do not have the same importance in the determination of densities.

5.3.4 Computation steps for the generation of network-constrained accessible areas

As the distances between the nodes have already been calculated, it seems advantageous to use this table to compute the network-constrained accessible areas from the centroids or the activities.

5.3.4.1 Extraction of the nodes' geometries

During this step, all the nodes located in the box created according to the bandwidth length are selected.

5.3.4.2 Sorting of the nodes in a "upper" and "lower" dictionary

It is important to sort the nodes according to the distance necessary to reach them from a given centroid. Therefore, the distances to the nodes are accessed using the table storing the distances between the nodes, and two dictionaries are created. During this step, one dictionary stores all the intersections (nodes) which are located further than the bandwidth, and the other dictionary stores all the nodes which are located closer than the bandwidth.

5.3.4.3 Comparison of the dictionaries and point interpolation

For each node in the "upper" dictionary, all the neighboring nodes (e.g. connected nodes) are selected using a dictionary created earlier. For each neighboring node falling in the "lower" dictionary, a point is created by selecting the corresponding segment and interpolating the remaining distance from the neighbor node to reach the bandwidth.

5.3.4.4 Convex Hull

Once all the points have been created and stored in a temporary list, they are used together with the nodes belonging to the "lower" dictionary to create a convex polygon whose geometry is stored in the database.

5.4 Original and visible segment approach comparison

Table 2 summarizes the main differences between both implementations.

Attributes	Original Algorithm	New Algorithm
Level of intermediate calculations (e.g. storage)	Low	High
Number of bandwidths at a time	One	Unlimited
Weighting factors	No	Yes
Possibility to select a certain kind of activity in the database	No	Yes
Performance for one computation	High	Medium to Low
Performance for many computations with different parameters	Low	High
Distances centroid-segment, event-segment	No	Yes
Closest segment approach	Yes	Yes
Visible segments approach	No	Yes
Storage of the accessible surface polygons	No	Yes
Centrality indexes	Yes	No

Table 2: Qualitative assessment of both algorithms

5.5 UML Diagram

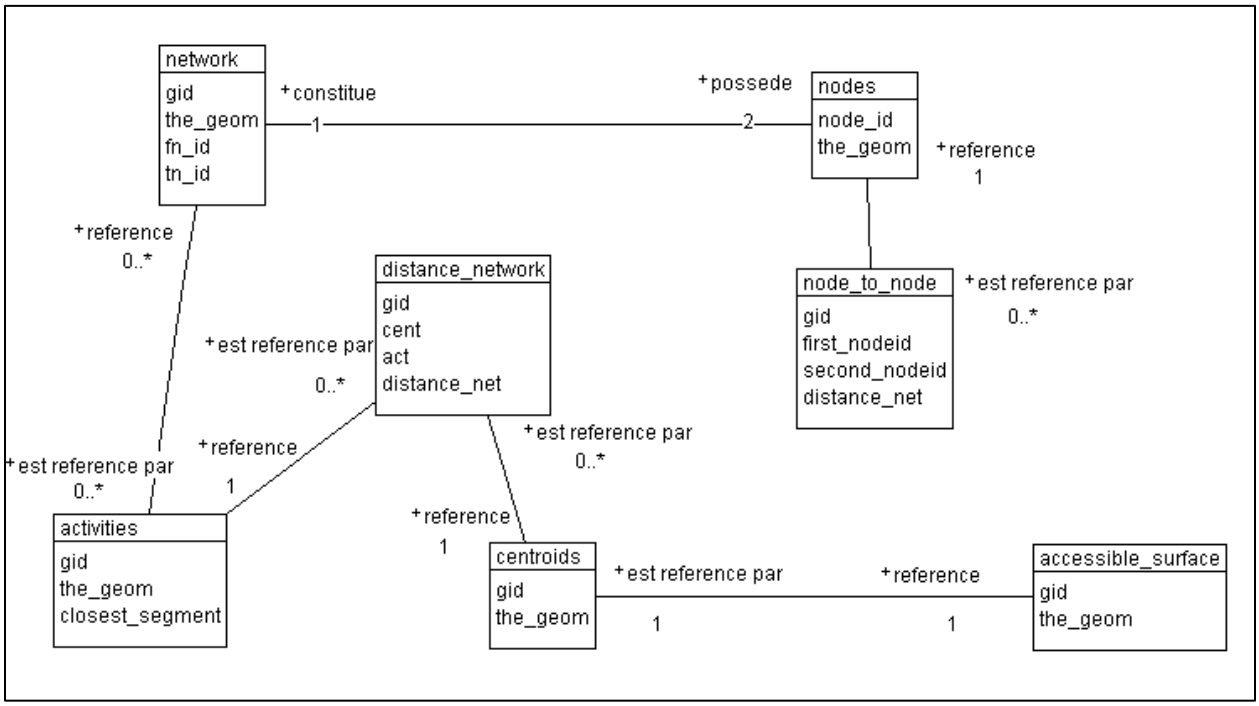

Figure 15: UML Diagram

5.6 Python PostgreSQL interactions

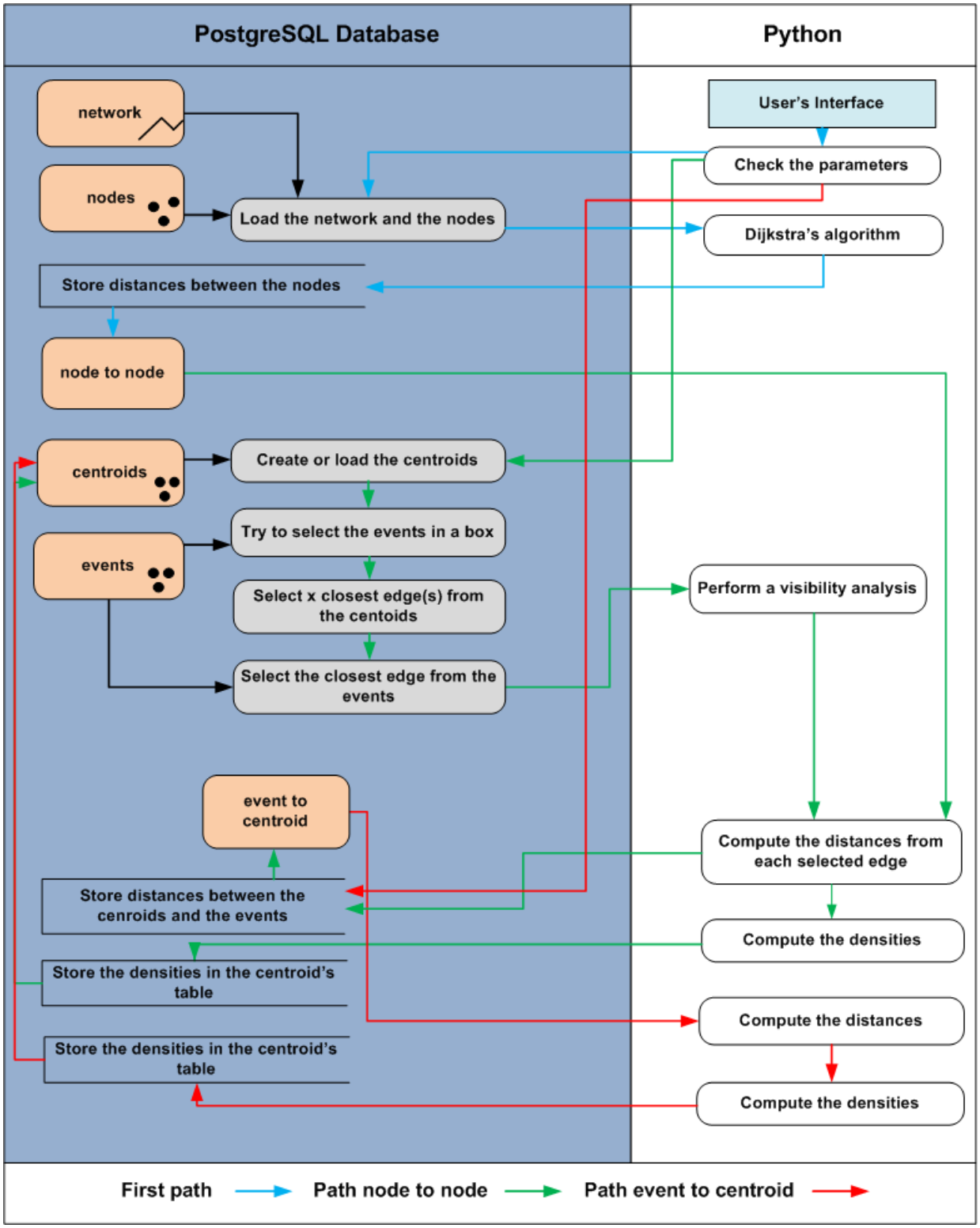

Figure 16: Fluxes diagram

6 KDE, "NetKDE closest" and "NetKDE visible"

The aim of this chapter is to understand the differences between the three approaches. First, an empiric analysis is carried out thanks to the visual analysis of densities computed on artificial networks. Secondly, as the differences between Euclidean distances and shortest-path distances measured on the road network are closely related to the densities computed with NetKDE, systematic ratios between these two distances are computed and analyzed for several cities. Finally, a case study using the three different methods is presented.

6.1 Simulations with artificial networks and activity configurations

When referring to the closest segment approach or NetKDE closest, one refers to the densities computed by selecting only the closest segment. When referring to the visible segments approach or NetKDE visible, however, one considers several closest segments and one performs the visibility analysis presented earlier.

6.1.1 Inputs

Table 3 summarizes the different simulations sorted out in ascending complexity to better understand the implications of each approach. All the maps are classified in 10 classes according to the centiles' density values.

Simulation Id	Network type	Spacing	Bandwidth	Grid resolution	Grid size	Events
1	Two segments	-	200 meters	1 meter	400x400	1 event in the middle
2	Regular squares	100 meters	200 meters	1 meter	600x400	1 event in the middle
3	Regular squares	25 meters	200 meters	1 meter	600x400	1 event in the middle
4	Regular squares	25 meters	100 meters	1 meter	600x400	10 events at random on the network

Table 3: Summary of the simulations

6.1.2 Empiric comparisons

6.1.2.1 Simulation 1

For the first simulation, only one event and two unconnected segments are used. When looking at the results for the closest segment approach on Figure 17 a), one can clearly distinguish the impact of projecting the centroids on one line only, as there is an obvious discontinuity located exactly between both segments. NetKDE visible (Figure 17 b) is more flexible, because the shortest-paths are calculated on both segments. This method provides a more continuous density surface in this case, but it is important to point out that it is an ideal situation, where the distance between both lines is exactly equal to the bandwidth.

Another interesting feature can be observed at the extremities of the segment around which the density values are distributed. As soon as the centroids' projection lines are no longer right-angled with the segment, the distances between the centroids and the segment are identical to the distances between the segment's node and the centroids. This artifact, if considered as such, could be corrected by using the Manhattan distance, which is equal to the root of the sum of the squares of the catheti.

6.1.2.2 Simulation 2

The simulations with a regular grid provide interesting results. When looking at Figure 17 a) for NetKDE closest, one can clearly distinguish on which segment the activity is projected. Within the same square defined by the network, densities suddenly change from locally high values to null values.

For NetKDE visible, the simulation is carried out twice, first five segments are considered, followed by eight. The impact on the resulting densities, when changing the number of selected line, demonstrates the importance of choosing the number of segments. When taking the 8 nearest segments (Figure 17 c) in the two squares defined by the network adjacent to the event, all the centroids within this space are projected on the segment where the event is located, which does not appear to be the case when selecting only the 5 closest segments (Figure 17 b).

When taking into account the 5 closest lines, and getting closer to an intersection, the densities are calculated on the basis of the 4 segments belonging to the intersection plus another one. Thus, all the visible lines within the square cannot be taken into account. With a regular gird, in order to make sure all the visible segments are selected, it is necessary to pick at least the 8 nearest lines.

The main differences between NetKDE closest and NetKDE visible appear at the vicinity of the event. Densities computed with the visible segment approach are always higher or equal to the ones obtained with the closest segment approach.

6.1.2.3 Simulation 3

This simulation is performed with the same parameters as the previous one, but with a denser grid. When only selecting the closest segment, density values are more irregular through Euclidean space. One can see how densities are strongly linked to the network. Taking all the possible visible segments into account is a much weaker hypothesis, because projecting the centroids on the closest line virtually divides Euclidean space according to the network configuration. Therefore, smoother density values can be obtained if more segments are taken into account. On Figure 17 a), due to the sudden change of projection line, one can observe that the densities within the squares of the grid are divided into two areas.

Results with a denser grid are nevertheless smoother for NetKDE closest because Euclidean space is divided into more subsets. This impact is less obvious for NetKDE visible, which is less sensitive to looser networks.

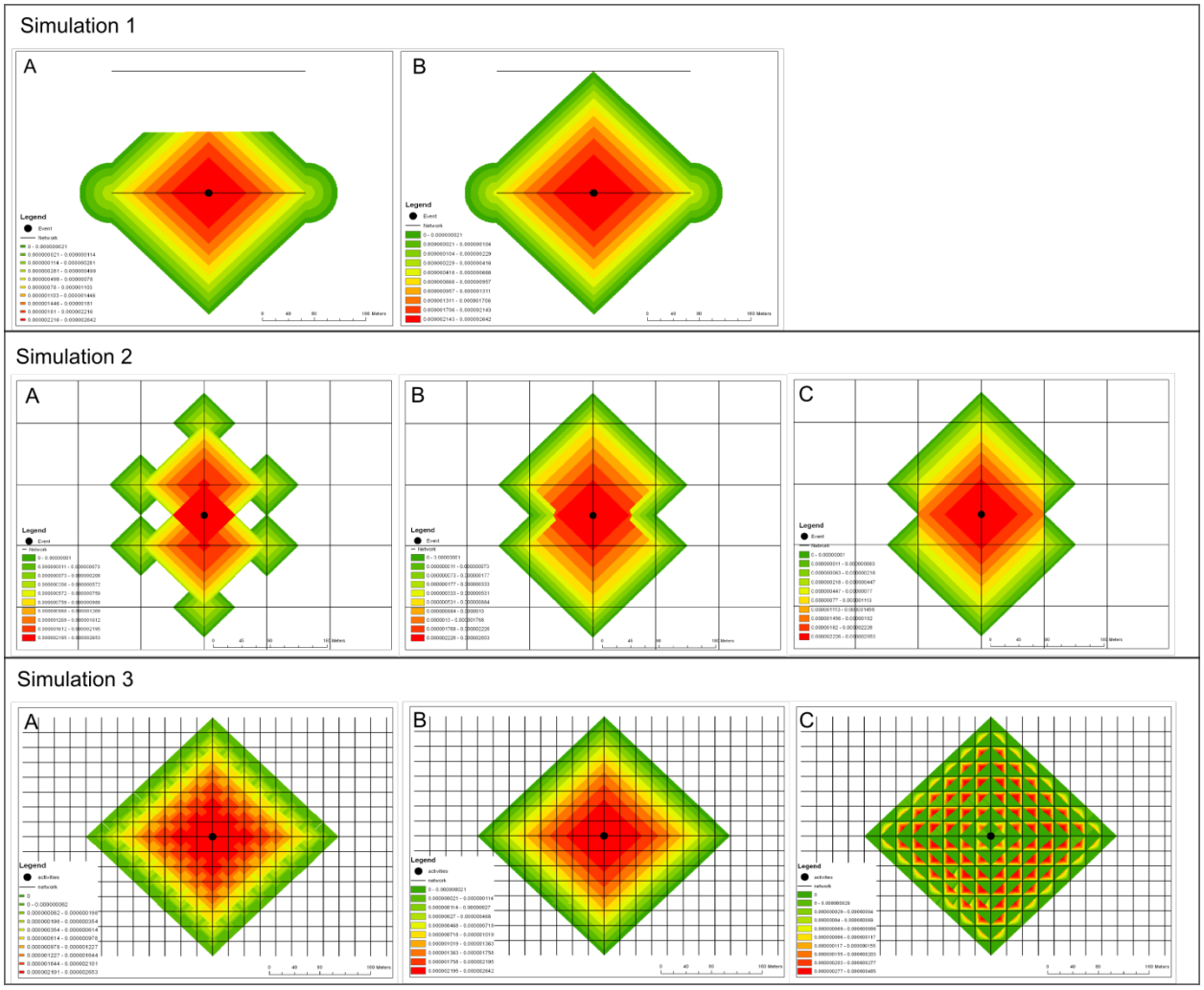

Figure 17: Simulation 1 a) closest segment approach, b) visible segments approach, Simulation 2 a) closest segment approach, b) 5 visible segments, c) 8 visible segments, Simulation 3 a) closest segment approach, b) visible segments approach, c) difference between b and a

6.1.2.4 Simulation 4

For this simulation, ten activities were generated at random on the network. As all the segments possess the same length and have the same weight, ten segments were selected at random, and then a number between 0 and 1 was generated randomly ten times. The lengths of the selected segments were multiplied with these numbers and the resulting distances were then interpolated on the lines to create the activities coordinates.

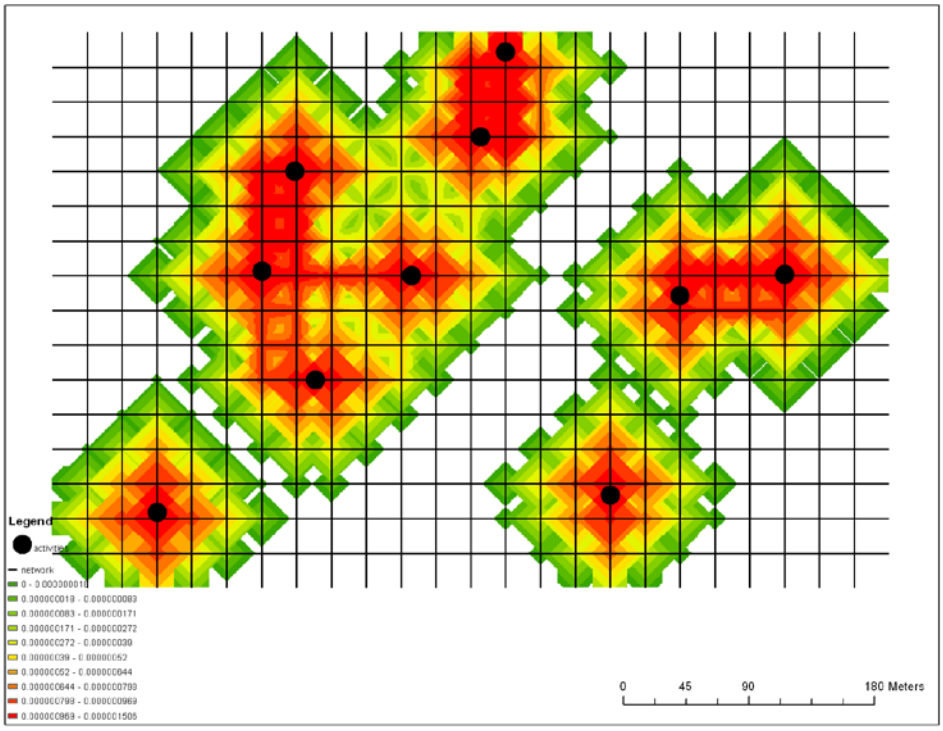

Figure 18: Simulation 4 closest edge approach

Once again, on Figure 18, one can see the divisions generated with such a strong hypothesis to implement a network constrained KDE. For this simulation, a conventional KDE was also performed on the dataset. When looking successively at the resulting density values of the three approaches, one can see the implications of decreasing the constraints applied to Euclidean space. When working in an urban environment, the network will always show a greater variability than the ones presented so far. As a result, more significant discontinuities might appear when selecting only the closest segment. The new approach developed in this work could be the optimal tradeoff between a NetKDE selecting only the closest segment, and therefore, producing rough densities and irregularities due to the artificial segmentation of Euclidean space, and the KDE which do not take into account the constraints imposed by the network. Computing density values for the entire space implies a combination of distances measured on the network and Euclidean distances. While NetKDE closest minimizes the distances measured in Euclidean space, NetKDE visible minimizes the sum of both distances, which translate the idea of an increased porosity between Euclidean space and the network space.

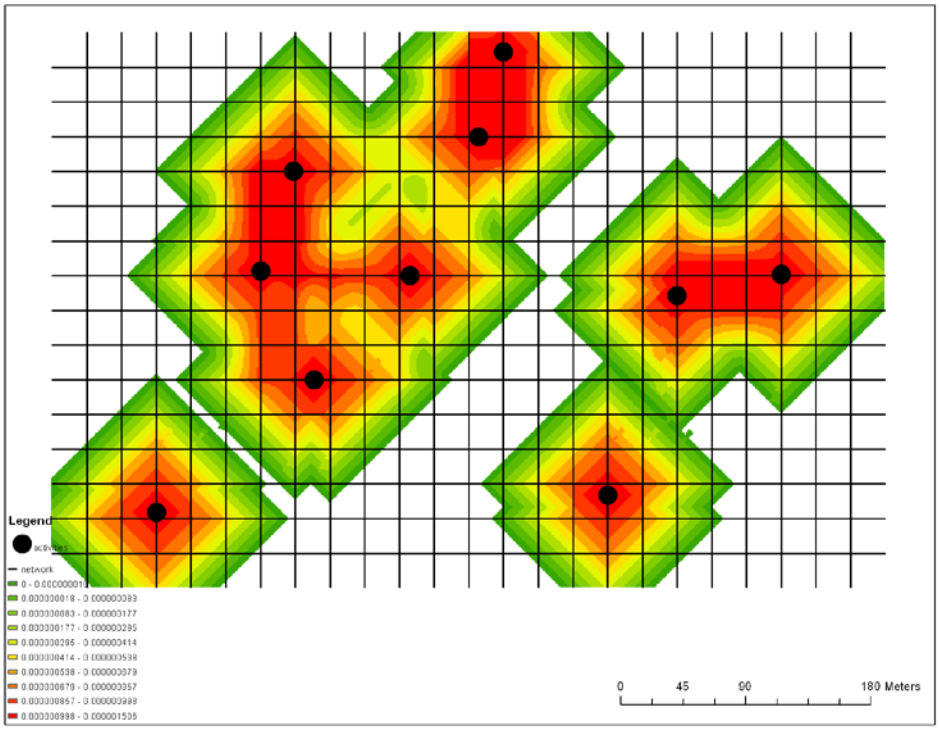

Figure 19: Simulation 4 visible approach with 8 edges

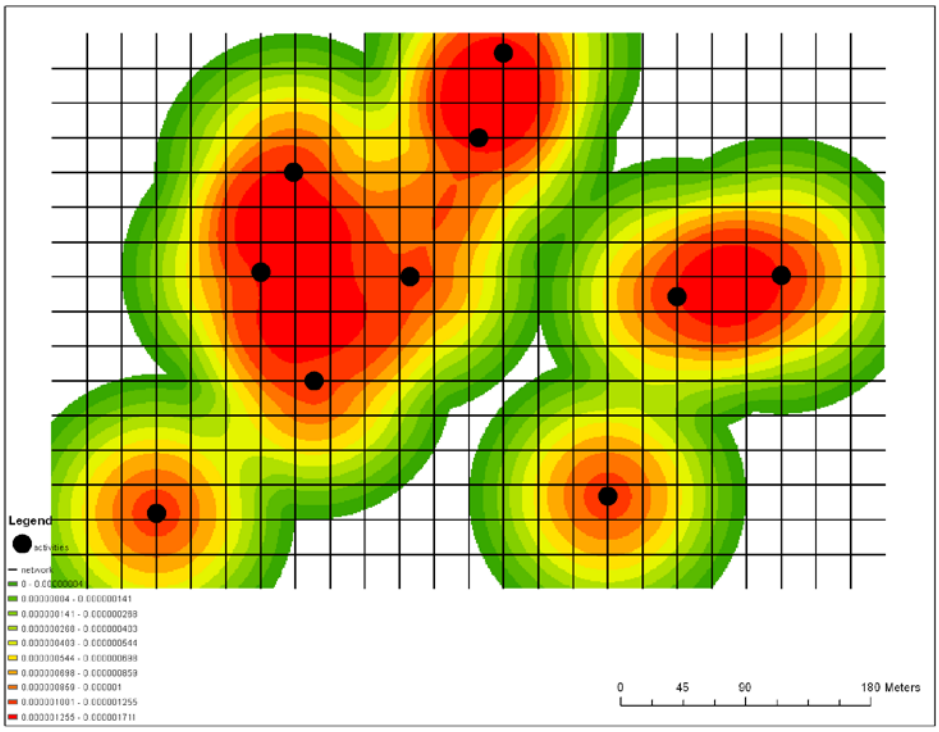

Figure 20: Simulation 4 with KDE

6.2 Ratio Euclidean distance, Shortest-path distances

Maki and Okabe (2005) have demonstrated that for a suburb of Tokyo the shortest-path distances were significantly different from the Euclidean distances, for Euclidean distances smaller than 500 meters. Differences between KDE and NetKDE densities are directly related to this phenomenon. Thus, in order to validate this first empirical approach, ratios between Euclidean distances and shortest-path distances are computed between the nodes of several city road networks. The goal of this experiment is to firstly, determine if the cities' network show the same characteristics or if there are significant differences among them, and secondly to assess the average difference between shortest-paths and Euclidean distances along the bandwidth gradient .

6.2.1 Inputs

The ratios are computed by taking into account for each node, all the other nodes within a Euclidean distance of 2000 meters. The networks of four cities are analyzed and presented in Table 4. For Geneva, the network of the whole canton is taken into account. However, the roads pedestrians could not access have been removed.

City	Number of edges	Number of nodes
Ljubljana	7'890	5'625
Geneva	10'860	7'471
Baghdad	66'648	43'856
Barcelona	11'180	6'485

Table 4: Cities network description

As millions of ratios are calculated for each city, the results are averaged and errors bars representing the standard deviations are provided. The histogram of Figure 21 shows on average how many times distances measured on the network are higher than Euclidean distances. Intervals always take into account the entire range of Euclidean distance starting from zero meters until the upper limit. This choice can be explained by the fact that when calculating density estimates, the entire range of distances is taken into account until the upper limit, represented by the bandwidth, is reached.

6.2.2 Results

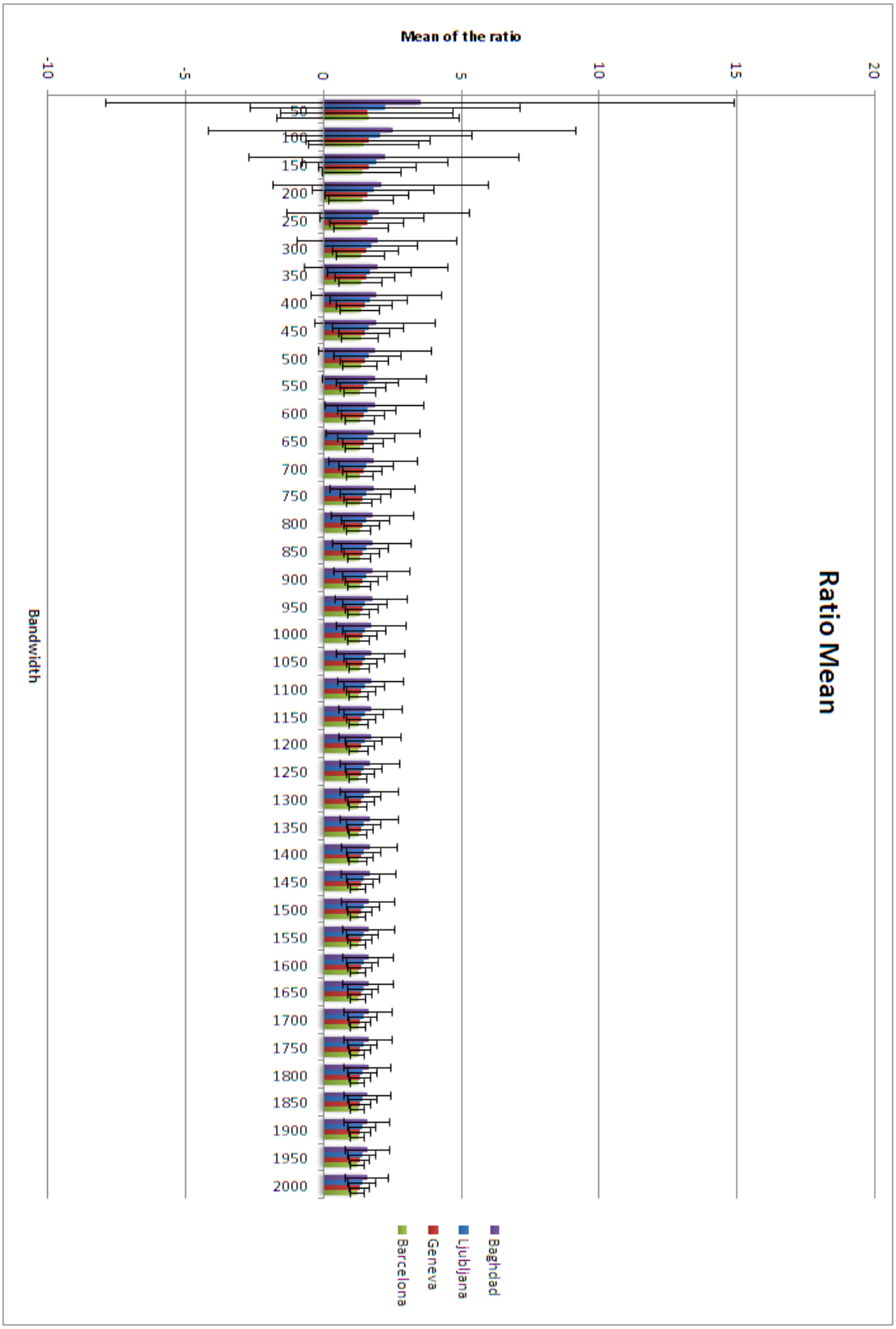
Mean of the ratio
-10
-5
0
5
10
15
20
Bandwidth
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
1150
1200
1250
1300
1350
1400
1450
1500
1550
1600
1650
1700
1750
1800
1850
1900
1950
2000
Ratio Mean
Barcelona
Geneva
Ljubljana
Baghdad

Figure 21: Ratios' mean for the cities of Ljubljana, Geneva, Baghdad and Barcelona

When studying Figure 21, one can notice that distances measured on the network differ more significatively from Euclidean distances between 0 and 500 meters. Indeed, it seems that the standard deviation of the ratios slowly reduce along the bandwidth gradient. For Euclidean distances between 0 and 100 meters, for Baghdad, Ljubljana, Geneva and Barcelona, the ratios' mean are respectively equal to 3.54, 2.03, 1.64 and 1.47. Thus, for a city like Baghdad, the difference between densities computed with NetKDE and KDE will probably be more significant than for a better connected city like Barcelona. At 500 meters, the ratios' mean seem to stabilize to values ranging from 1.34 to 1.62. For a bandwidth of 2 kilometers, the ratios' mean oscillates between 1.25 for Barcelona and 1.6 for Baghdad. Hence, the differences between Euclidean distances and shortest-paths are negligible for high bandwidths, and the assumption of walking is no longer a reasonable hypothesis. Therefore distances must be transformed in time to take into account the impact of the road network constraints. As a result, NetKDE evaluated in cities should only be considered for relatively short distances, where the assumption of walking is still valid. Due to the conjunction of the road network constraints and the limited differences between Euclidean distances and shortest-paths for high bandwidths, one should consider that using KDE in these cases is a reasonable approximation of NetKDE, which is both easier to implement and less time consuming.

6.3 KDE, NetKDE closest and visible on the retail stores in the city of Geneva

The aim of this section is to present the results of the three approaches for a real case-study. The resulting density surfaces are compared amongst each other, and these results are linked to the ones obtained in the previous sections.

6.3.1 Methodology

Three different simulations were performed on the entire city of Geneva on the NOGA category "retail stores". Overall, there are 78'792 centroids with a spatial resolution of 25 meters. Successive bandwidths starting from 100 meters until 1'000 meters have been evaluated for each method. The densities have been calculated for each approach on the basis of the projection of the retail stores on their closest segment. For "NetKDE closest" and "NetKDE visible", the distance between the centroid and its projection on a segment was taken into account. The 10 closest segments were selected in order to perform "NetKDE visible". By using the table where the distances between the nodes are stored, the computing time to calculate the densities for all the bandwidths is equal to about 10 hours for "NetKDE closest" and 28 hours for "NetKDE visible".

In order to evaluate the difference between each method at different bandwidths, the standard deviation between the three couples of maps is computed. Therefore, the equation used is

$$STD(KDE, NetKDE) = \sqrt{\sum_{i=1}^{n} (KDE_i - NetKDE_i)^2 / n}$$

where KDE_i and $NetKDE_i$ (closest or visible) are the density values of the ith cell, and n the total number of cells (or centroids) in the study area.

6.3.2 Results

The maps are presented in Figure 50, Figure 51 and Figure 52 of the appendix. The differences between the three approaches are highlighted on Figure 22 below. First, it should be noted that the standard deviation is very high compared to the maximum values for "NetKDE closest" and "NetKDE visible", which are respectively equal to $4.3*10^{-9}$ and $5.9*10^{-9}$ at 500 m. "NetKDE closest" possesses a higher standard deviation for all the bandwidths when compared to KDE. Similarly to the ratios computed in the previous section, the variability decreases along the bandwidth gradient. It seems that this variability tends to stabilize for higher bandwidths. When comparing "NetKDE visible" and KDE, one can see that the maximum standard deviation is not found for the smallest window width. The variability increases until 300 m and then decreases slowly until 1'000 meters.

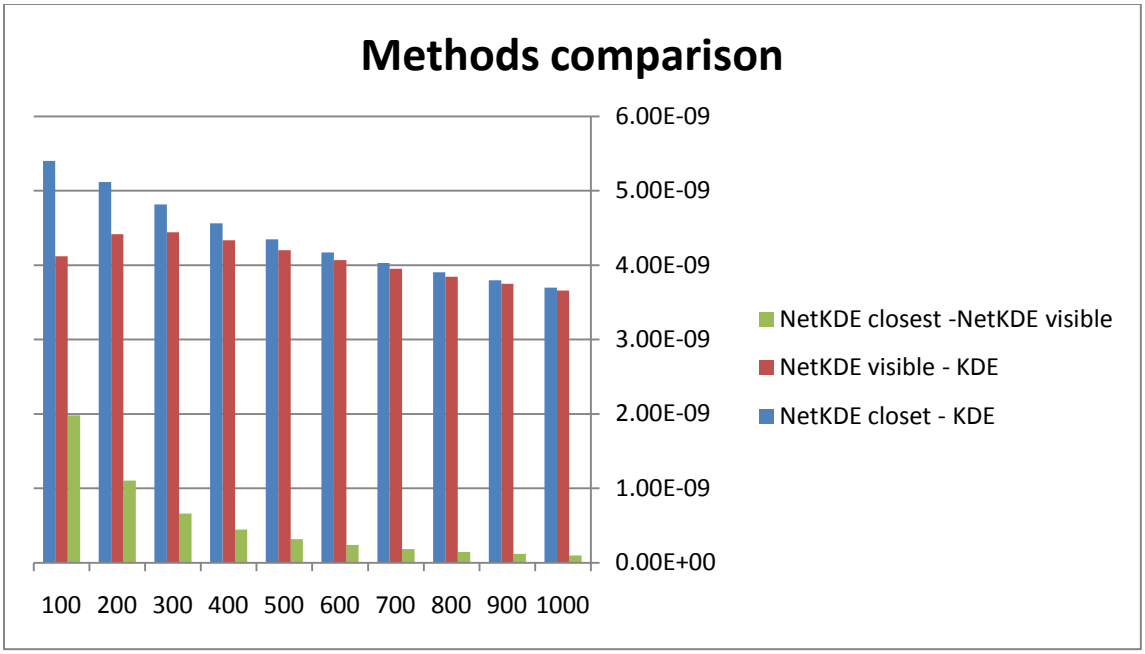

Figure 22: Methods comparison

Theses simulations confirm the observations made in the previous chapter. "NetKDE visible" is at the interface between KDE and "NetKDE closest". Nevertheless, the densities generated with "NetKDE visible" seem more like the ones generated with "NetKDE closest", than the ones generated with KDE.

The closest segment approach is more sensitive to the location of the centroids than the visible segment approach. When comparing Figure 50 and Figure 51, "NetKDE closest" exhibits more discontinuities than "NetKDE visible" and is, therefore, more difficult to interpret. By choosing a fine resolution grid and by using "NetKDE visible", it is possible to create a smooth density surface which takes into account the network morphology constraints. Table 5 below summarizes the differences between the three approaches.

	NetKDE closest	NetKDE visible	KDE
Network constrained	Yes, minimizes the distances in Euclidean space	Yes, minimizes the sum of the distances measured in Euclidean space and on the network	No
Absolute density values	Low	Medium	High
Sensitivity of the results according to the centroids locations	High	Medium	Low
Smoothness of the density surface	Low	Medium	High
Computing time	Medium	High	Low
Relevance in urban context	Medium	High	Medium - Low
Relevance along a linear network	High	Low	Low

Table 5: Summary of the differences between NetKDE closest, NetKDE visible and KDE

7 A building based neighborhood analysis of economical activities in Geneva

This section presents an approach where the network based indicators presented earlier are calculated from each building in the canton of Geneva. First, the advantages and implications of such an approach will be discussed. Secondly, the results will be presented, and thirdly, they will be discussed and the future prospects of this work will be presented.

7.1 Introduction

In this study, the approach implemented is slightly different from the ones presented earlier. Instead of using a regular grid through Euclidean space, the network based density values are evaluated on a building basis. The initial motivation for performing such a simulation is based on the idea that if displacements within a city are constrained by the network, starting points and destinations are then, in most cases, constrained by the built areas. Moreover, there are a significant number of advantages for carrying out such an analysis. Firstly, the ambiguity discussed earlier arising from the combination of Euclidean and network based distances, when computing NetKDE for the entire study area, is addressed by measuring distances on the network only. In addition, considering that the locations of the buildings are constrained by the network, and consequently the locations of the economical activities as well, one can make the assumption that within a city each building can be connected to a main access road. Secondly, the number of centroids treated for such a configuration may drop considerably and, therefore, the computing time as well. Thirdly, if density values of events are calculated only on a building basis, it clears a significant surface of the area under study enabling a better visualization of the results, which makes it possible to locate objects of interest in the city and gather accurate information about the relative intensity of a type of economic activity in the direct neighborhood. Fourthly, the initial buildings' attributes can be crossed with any indices or densities of interest. Thus, the network based indices and densities can be considered as an attribute of the building's neighborhood and could be used to point out the interactions between spatial actors and residents for instance.

7.2 Methodology

The city of Geneva was selected in order to perform this analysis. The study initially requires three main files. All the data used in this case study came from the SITG ("Système d'information du territoire genevois"). As the analysis focuses on short distances and is an attempt to characterize the density of activities in the direct neighborhood of a building, and because integrating vehicles in the calculations considerably increase the complexity of the model, only the roads accessible to pedestrians were taken into account. The network's topology was then cleaned and each node was given a unique identifier. The centroids were generated with ArcMap on the basis of the buildings' footprints. In total, densities have been computed for 13'714 buildings. The density values were first calculated on the centroids and then transferred to the buildings for the visualization. The approach used for this simulation only considers the nearest segment from a given centroid.

The activities analyzed are classified according to the NOGA classification. Following a preliminary simulation on the entire dataset to store the distances between the centroids and the activities, several subsequent simulations were performed to determine the densities of different types of activities. The density values were mainly calculated on the second level of the NOGA ("nomenclature générale des activités économiques") classification while the diversities were computed on the third level. The details of the classifications can be found in the appendix.

A bandwidth of 500 meters was selected in respect with city's human scale (walking space) (Produit, Lachance-Bernard, Strano, Porta, & Joost, 2010). Moreover, as shown earlier, the differences between Euclidean distances and distances measured on the network are more significant for Euclidean distances shorter than 500 meters. In addition, in order to avoid any biases due to the edge effect, densities were computed for the entire canton of Geneva, but only the results of the city are presented in this chapter.

Firstly, the overall accessibility of the city is presented as the relative accessible surface from a given building and the number of intersections falling within the generated polygons. The entropy and the dominant species are presented on the second level of the NOGA classification. The categories selected to compute the diversity indices are the same than the ones used to compute the densities.

Secondly, the densities of the twelve most common types of economical activities are calculated. Table 6 shows the activities selected for the simulation. A Simple K means clustering method is then applied on the densities obtained for each building. Although this method is inherently aspatial, because it only takes into account the values of the buildings attributes, the spatial dimension is indirectly introduced in the model thanks to the density values that are defined in space.

Type of economical activity	NOGA number	Example of activities	Total number of activities
Retail stores	47	Supermarket, bakery, butchery	2'885
Catering	56	Restaurant, bar, dancing	1'544
Wholesalers	46	Wholesalers for fruits, plants, vegetable, clothes	1'530
Activities for human health	86	Hospital, dentist, medical laboratories	1'519
Activities of the social head offices; management advices	70	Public relations counsel, societies social head office	1'221
Legal and accounting activities	69	Law firms, accounting and trusties activities	1'129
Auxiliary activities for financial services and insurances	66	Administration of financial markets, brokerage	1'092
Financial services	64	Private bank, national bank, cantonal bank	957
Other personal services	96	Laundry, barbershop	914
Associative organizations	94	Labor union, religious associations, political organizations	890
Specialized construction work	43	Demolition, installation of sanitary facilities	650
Real-estate activities	68	Real-estate agencies, rental of land	624

Table 6: Types of economical activities selected

7.3 Results

In this section, the accessible surface index and its application to measuring diversities of economic activities will be described first. Then, the clusters generated on the densities of economic activities will be presented and commented on, together with the indices introduced in the previous sub-section.

7.3.1 Accessible Surface Index

As presented earlier, instead of using a disk to determine the potential accessible area from a given centroid, one can use the network and a convex hull function. The aim of this section is to provide a relative accessibility index based on the method described above. In Figure 23, one can appreciate various kinds of polygons for different bandwidths and network configurations.

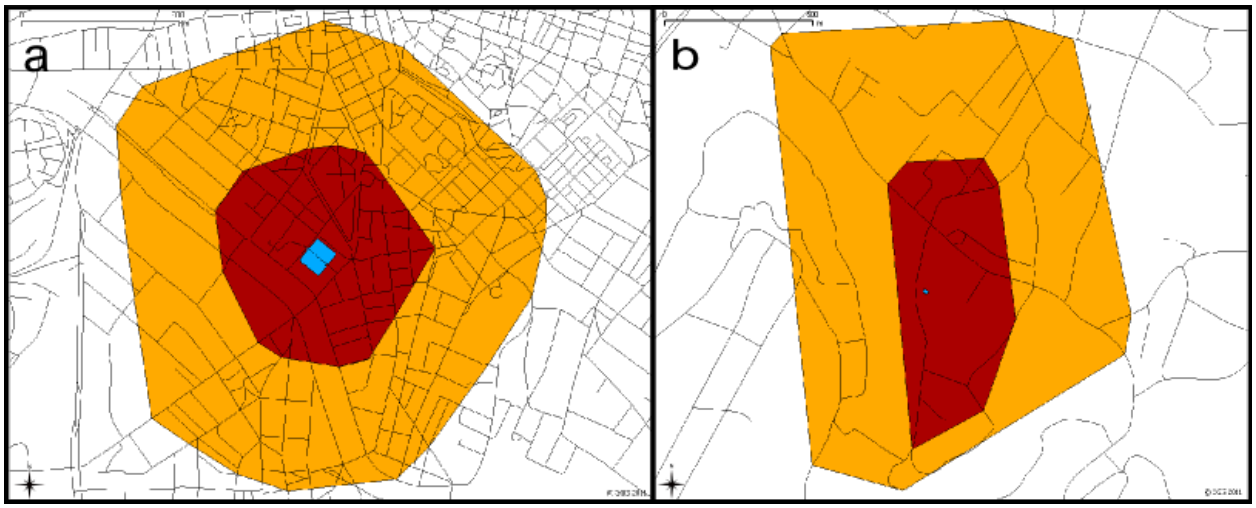

Figure 23: Differences in the search areas between two buildings in Geneva. The area in orange represents the resulting polygon for a distance of 1000 m on the network, while the red polygon is the result of a search area of 500 m. a) The building is located where the network is well connected (scale bar = 600 m) b) the building is located in an area where the network is less dense (scale bar = 500 m)

In order to facilitate the comprehension of the legend and provide a better insight to the meaning of this index, the surfaces of the polygons were normalized with the corresponding surface in Euclidean space (e.g. a circle), then multiplied by 100. Thus, the values range theoretically from 0 to 100% and represent how close the polygons are from a circle at a given scale. The results of this computation with a bandwidth of 500 meters are presented on Figure 24. The classes have been generated using the deciles of the dataset. The first observation one can make, is that in enclosed areas like in box 1 at “La Jonction”, the relative accessible area is much more limited than in the city center for instance.

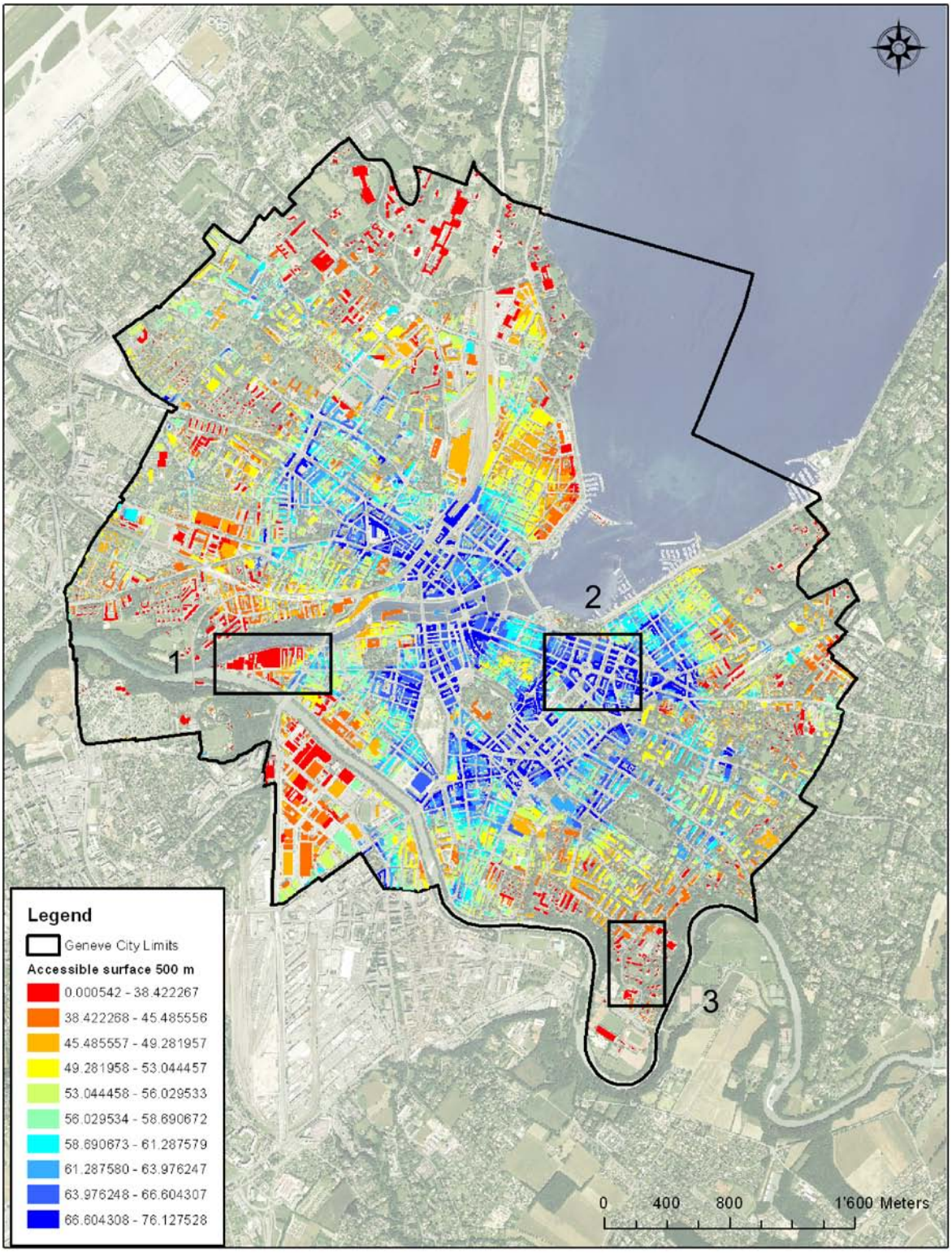

Figure 24: Accessible surface using a bandwidth of 500 meters, 1) La Jonction, 2) Rond-Point de Rive, 3) Le Bout-du-Monde

Figure 25: Zoom on the buildings and comparison between a good access index on the left (box 2) and a bad access index on the right (box 3)

One can see on Figure 25 (boxes 2 and 3), the importance of the network connectedness and density in evaluating the accessible surface.

As a complementary study, the number of nodes falling within the accessible surface of each building has been computed in an attempt to determine the relationship between the number of intersection and the surface of the polygons. The comparison between both indicators is presented in Figure 26.

An exponential curve was fitted to the dataset. The resulting standard deviation between the sample points and the curve is equal to 26.35. The variability of the number of nodes falling in the accessible area between 60 and 75% is very high. This indicates that, even if these two concepts seem closely related, a high potential accessible area does not guarantee a high network connectedness. Thus, while the potential accessible area is a measure of how far one can go from a starting point in every direction, the number of intersections better represents the idea of how well one can travel within this space, and is therefore, a direct measure of network connectedness. Connectivity measures the directness of the pathway between households, shops and places of employment and is based on the design of the street network (Leslie, Coffee, Lawrence, Owen, Bauman, & Graeme, 2007). By first measuring the potential accessible area, one can take into account the barriers of the built environment, and then measure the number of options for the travel routes within this space.

When comparing the maps for both indicators, the map representing the number of intersections falling within each polygon seems more clustered than the original one. As previously introduced in this work, the density of junctions has already been used in previous studies to determine the network density in the city of Trieste in Italy (Borruso G. , 2003). This method could be considered as a similar approach used to find clusters of different network subsets in a city. Therefore, Figure 27 provides a qualitative measure of where pedestrians are favored the most in Geneva. The street connectivity is also measured in the context of walkability studies by counting the number of intersections in predefined grid of points (Leslie, Coffee, Lawrence, Owen, Bauman, & Graeme, 2007). Therefore, the technique introduced in this section to measure street connectivity is an adaptation of existing measures to the "building model".

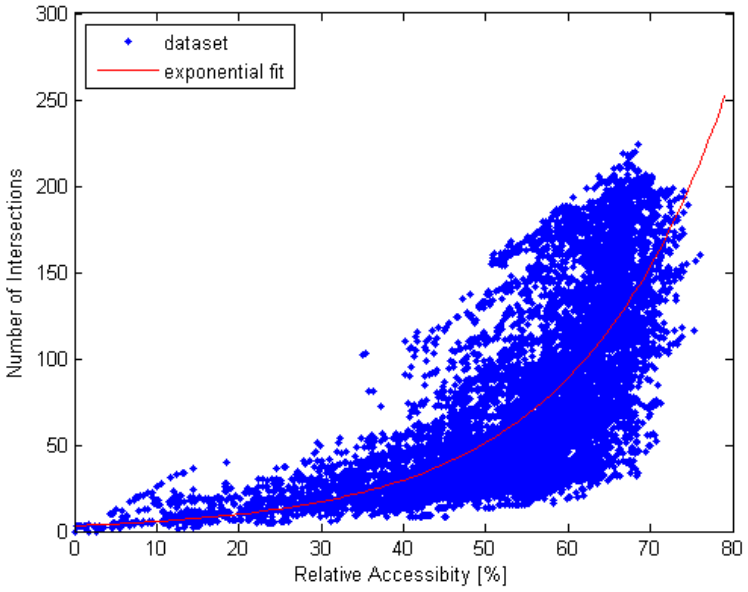

Figure 26: Relationship between the areas of the polygons and the number of intersections falling within them, an exponential curve was fitted on the dataset (Equation = 3.31*EXP (0.0548*Relative Accessibility with STD = 26.35)

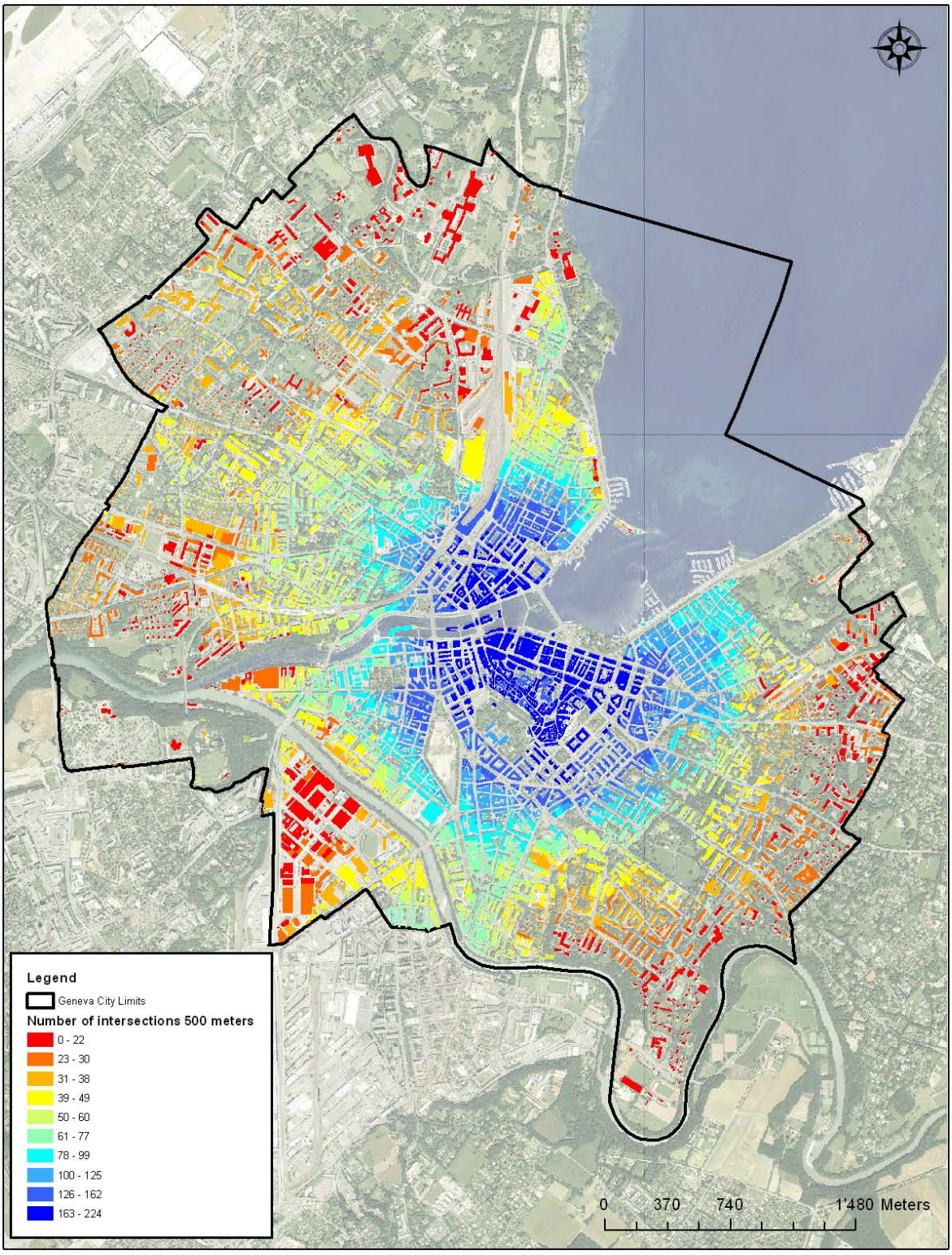

Figure 27: Number of intersections in each polygon (500 meters)

7.3.2 Diversity Indices

The polygons previously generated can now be used to study other characteristics of the city. By counting the number of points falling within the "accessible" areas, one can compute a network based diversity estimation. The diversity indices have been computed using the second level of the NOGA classification, for the twelve selected types of economic activities.

As a first general observation, it can be pointed out that the entropy (Figure 28) and the relative dominancy of species (Figure 30) are more or less inversely proportional. This makes sense when considering that the more strongly an activity is represented, the less important the overall entropy of the system will be. On Figure 29, the class "retail stores" dominates most of the city center. At the periphery, one can start distinguishing specialized sector where other types of activities dominate. The activities are sorted in decreasing number of occurrences.

By combining the three maps, one can represent the spatial distribution of economic activities in the city. On Figure 28 1), for instance, one can see that this district possess a very low entropy and is dominated at a significant level (Figure 30) by the class retail stores (Figure 29). When moving along the eastern side of this area, the entropy drops as does the relative dominance of this class. Therefore, one can conclude that this area is the heart of this class in Geneva.

In box 2 of Figure 28, the district of "Eau vives" possesses a high entropy index. The class "retail stores" is still the most common, but it represents approximately only 15 to 18% of the total of economic activities.

The district of "Les Pâquis" has relatively low entropy and the class "restoration" dominates (Figure 29 1) the other classes at a level of about 21%. This area is the only one close to the city center that seems to be really specialized in this sector.

In box 2 of Figure 29, one can see that this area is specialized in the health sector. Nevertheless, the entropy is still quite high in this district and the relative dominance of this class is quite heterogeneous.

In box 3 of Figure 30, the district of the associations (United Nations, World Health Organization …) can be clearly distinguished. The area is highly specialized, and, therefore the entropy is very low, except on the East side by the lake.

The analysis diversities maps is far from being completed, but the aim of this section is to show how one can compare these indexes to gather information about the distribution of economical activities in different parts of the city. These methods can then be implemented according to the user's needs by selecting desired parameters.

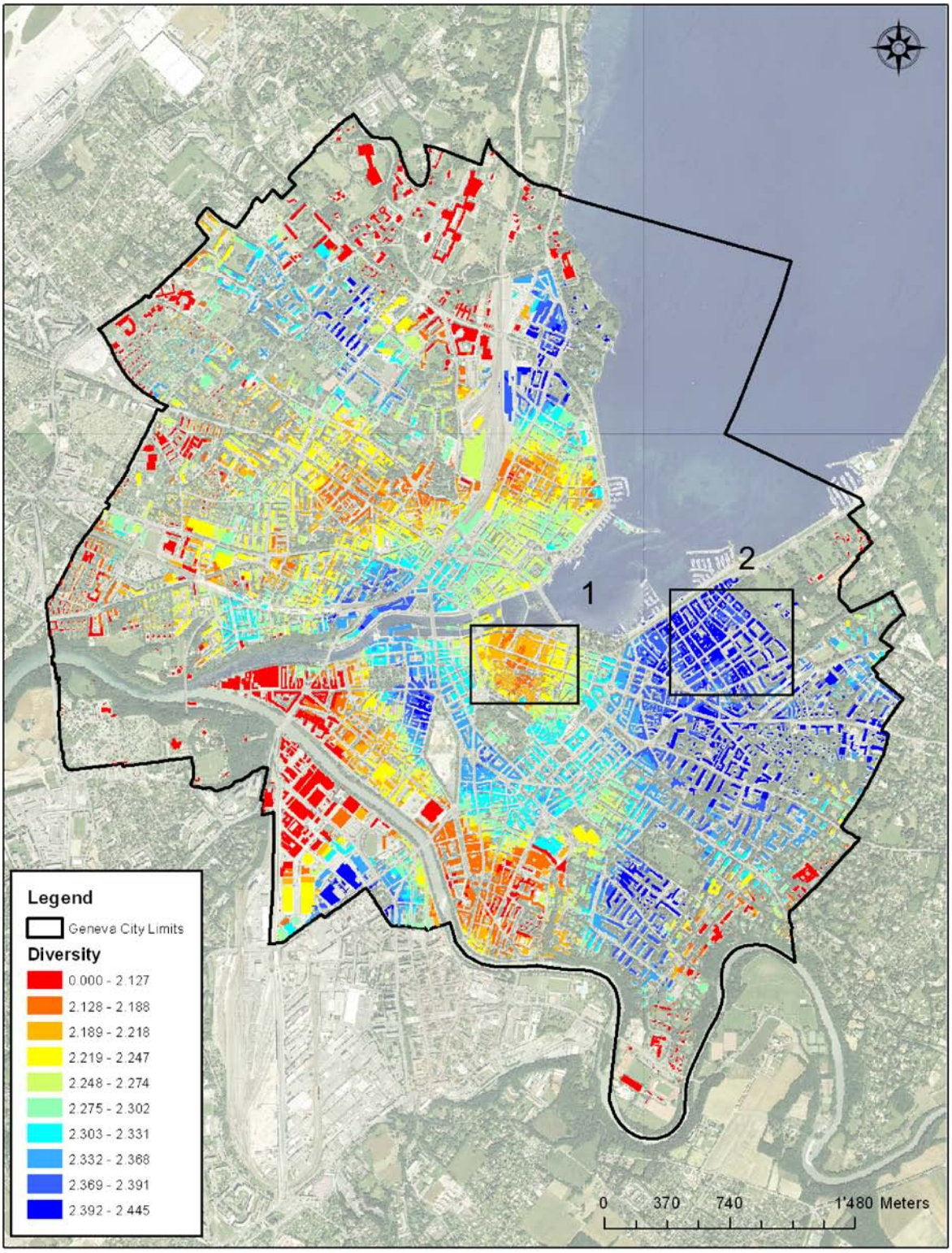

Figure 28: Diversity on the second level of NOGA 500 m, 1) Cité 2) Eaux-vives

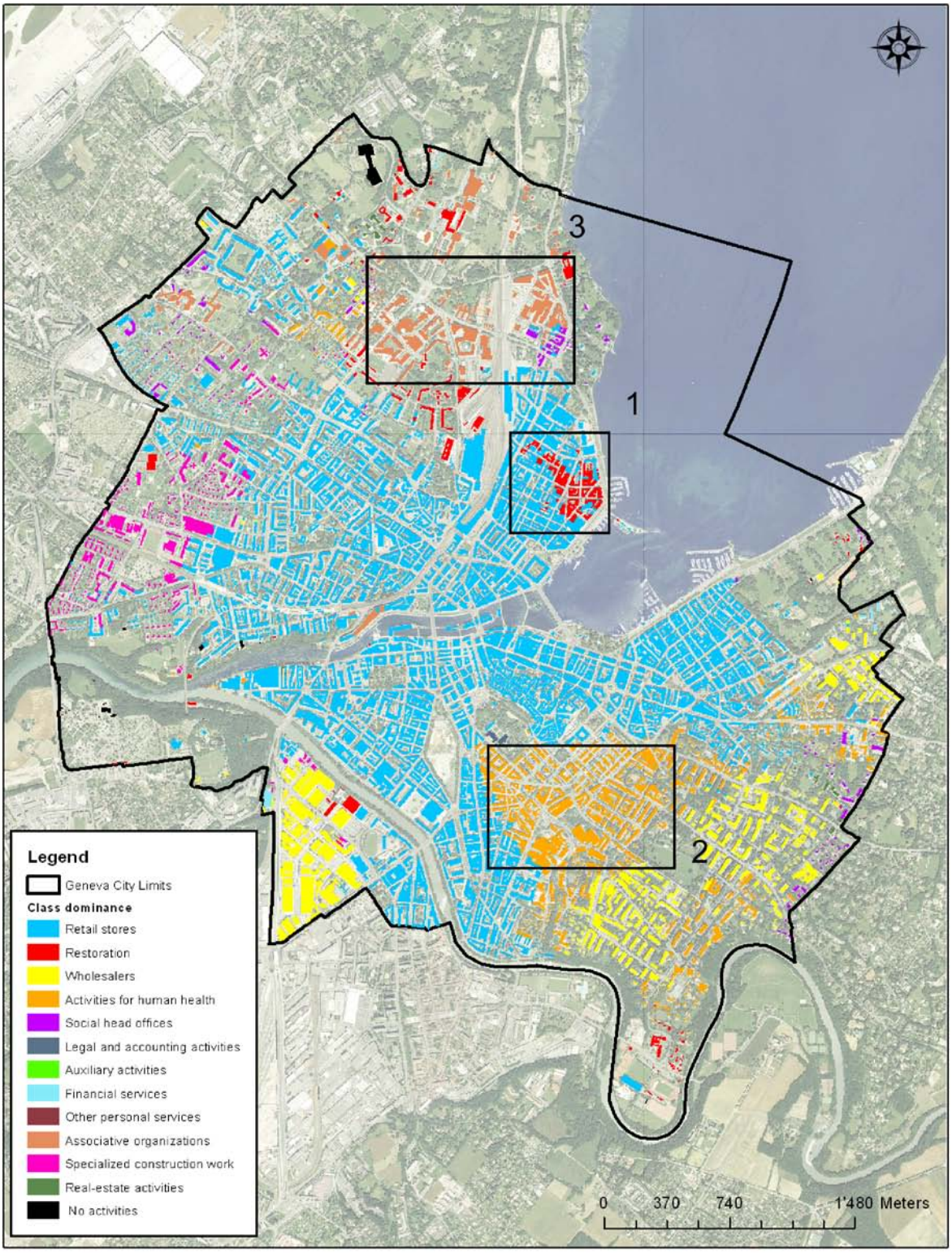

Figure 29: Class dominance on the second level of NOGA classification 1) Les Pâquis 2) Champel 3) Associations district

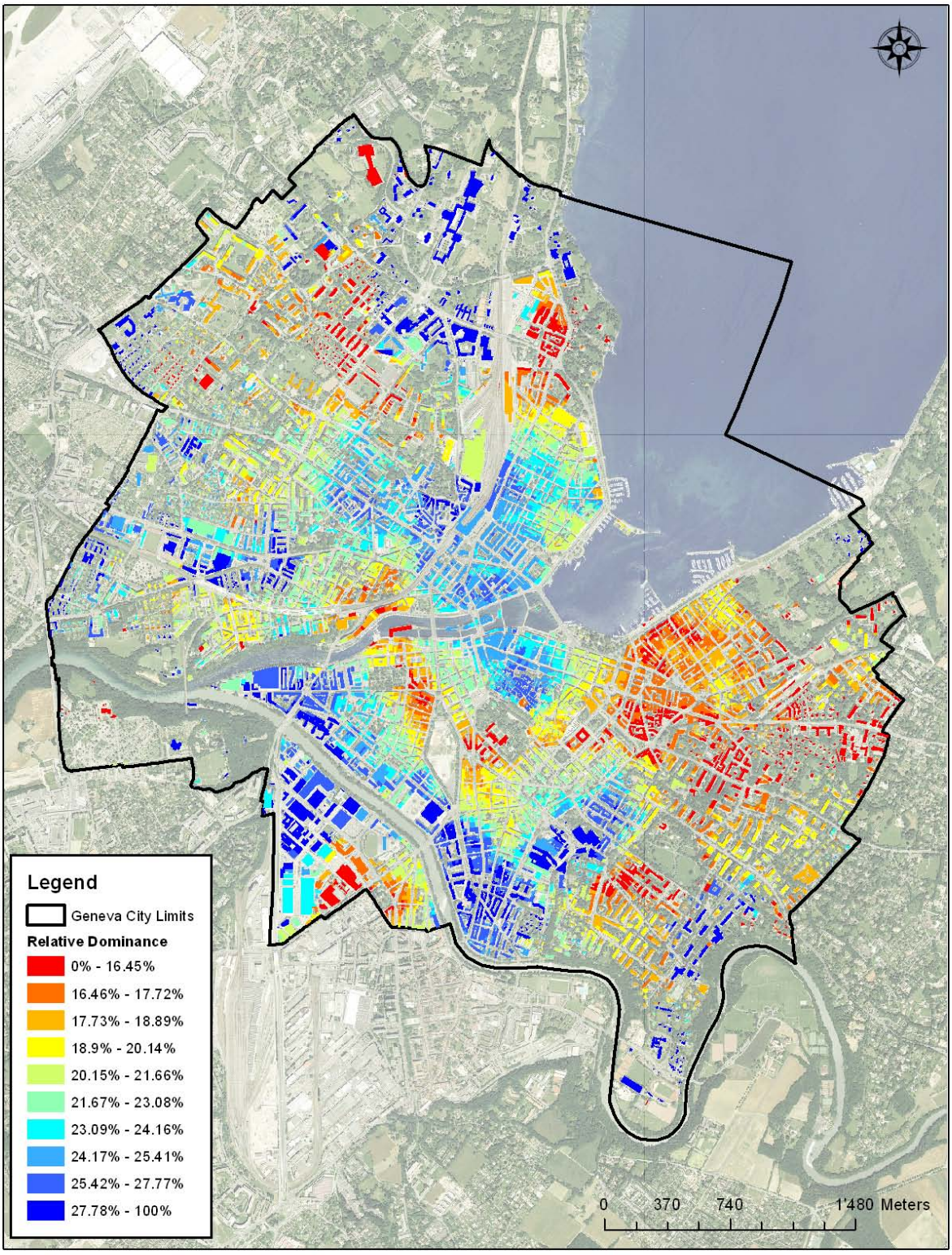

Figure 30: Relative Dominance of the second level of NOGA classification

7.3.3 Density indexes

The densities of the twelve most common types of economic activities in the city of Geneva are presented on Figure 53, Figure 54 and Figure 55 of the appendix. The scale of the densities is the same for every NetKDE simulations, and, therefore, they are comparable amongst each other. Overall, 4 different classes are generated. On Figure 31, one can distinguish in green the "**main economical center**" of Geneva for the selected types of activities. The highest density values of "wholesale", "real-estate activities", "financial services", "retail stores", "auxiliary services (financial and insurances)", "social head offices" and "legal and accounting activities" are all grouped in this category (Figure 32). Moreover, they all possess a very high correlation between each other (Figure 33). As a result, one can assume that these activities tend to attract each other. Some of the highest values for the activities of "specialized construction works" and "other personal services" can also be found in this category, but one can find a higher heterogeneity of categories within their range of high values. The entropy in this cluster is relatively low (Figure 28).

The "**secondary center**" appears as an extension of the main center, as all the preceding types of activities possess slightly lower density values. Classes that dominate this category are "activities for human health", "catering" and "associative organizations". Thus, compared to the main center, offices start to slowly disappear to the benefit of other services. This observation is confirmed by the fact that diversities are usually higher in this class (Figure 28). The class of "other personal services" is also well depicted in this category and is highly correlated to the class "catering". Both are also correlated to a slightly lower degree to the class of "retail stores" (Figure 33). Therefore, at the immediate neighborhood of the "main economical center", common services of everyday life become increasingly important. On the left side of the Rhône (Figure 31 3), one can see that the districts of "Saint-Gervaix" and "Genève-Cité", which are well known for their high densities of economic activities, appear in the two denser classes.

The "**transition zone**" possesses low density values of the most important classes present in the main economical center. In this cluster, one can find mainly intermediate values for the classes "other personal services", "specialized construction works", "restoration", "human health activities" and "associative organizations". On Figure 32, one can see that the "transition zone" possesses a wider range of density values for nearly all types of activities. As a result, this category has highly variable entropy indices. A more refined characterization of this zone could be performed by increasing the number of clusters.

Finally, the "**residential zone**" groups together most of the lowest density values of the entire data set. Classes such as "financial services", "real-estate activities" and "retail stores" disappeared from this cluster. The only types of economic activities retaining slightly higher density values are "other personal services", "specialized constructions" and "associative organizations". The district of "Les Tranchées" (Figure 31 2) is the main residential area in the center of Geneva.

When comparing the categories obtained with Figure 27, one can clearly see that the main pedestrian areas (above 78 intersections) match the “main economical center”, the “secondary center” and the “transition zone”. Expect for the category “residential zone”, all the other categories are under the domination of retail stores Figure 29. On Figure 31, one can also see that there is a cluster of “residential zone” in between the main economical center and the “secondary center”. This can be explained by the fact that the “Parc des Bastions” is separated from the historic district by the Treille’s wall (Figure 31 1). This therefore proof that where a classical KDE would fail to assess such a configuration, NetKDE can capture the underlying constraints of the network and, therefore, provide more accurate information about the phenomenon being studied.

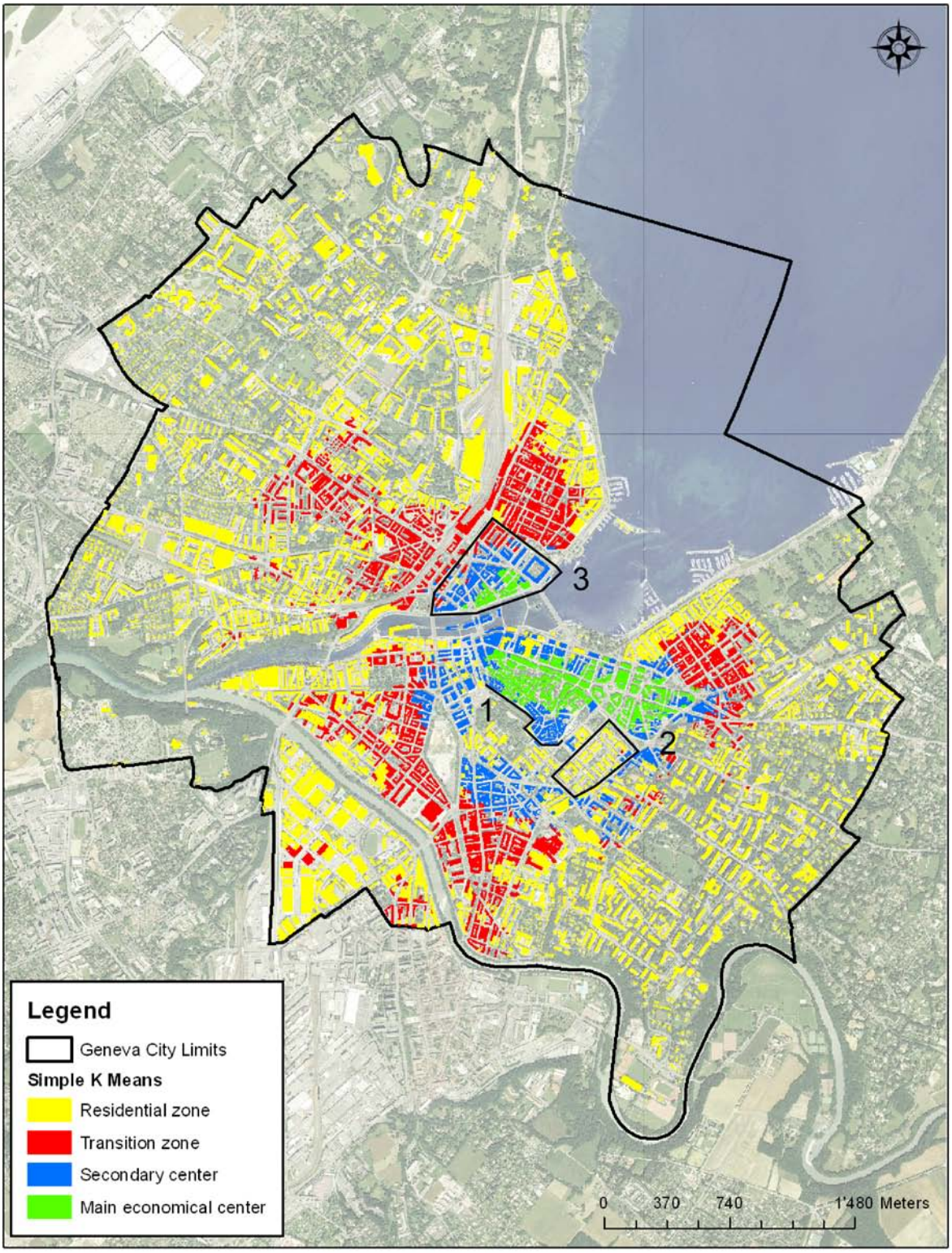

Figure 31: Simple K means clustering on economical activities densities with K = 4

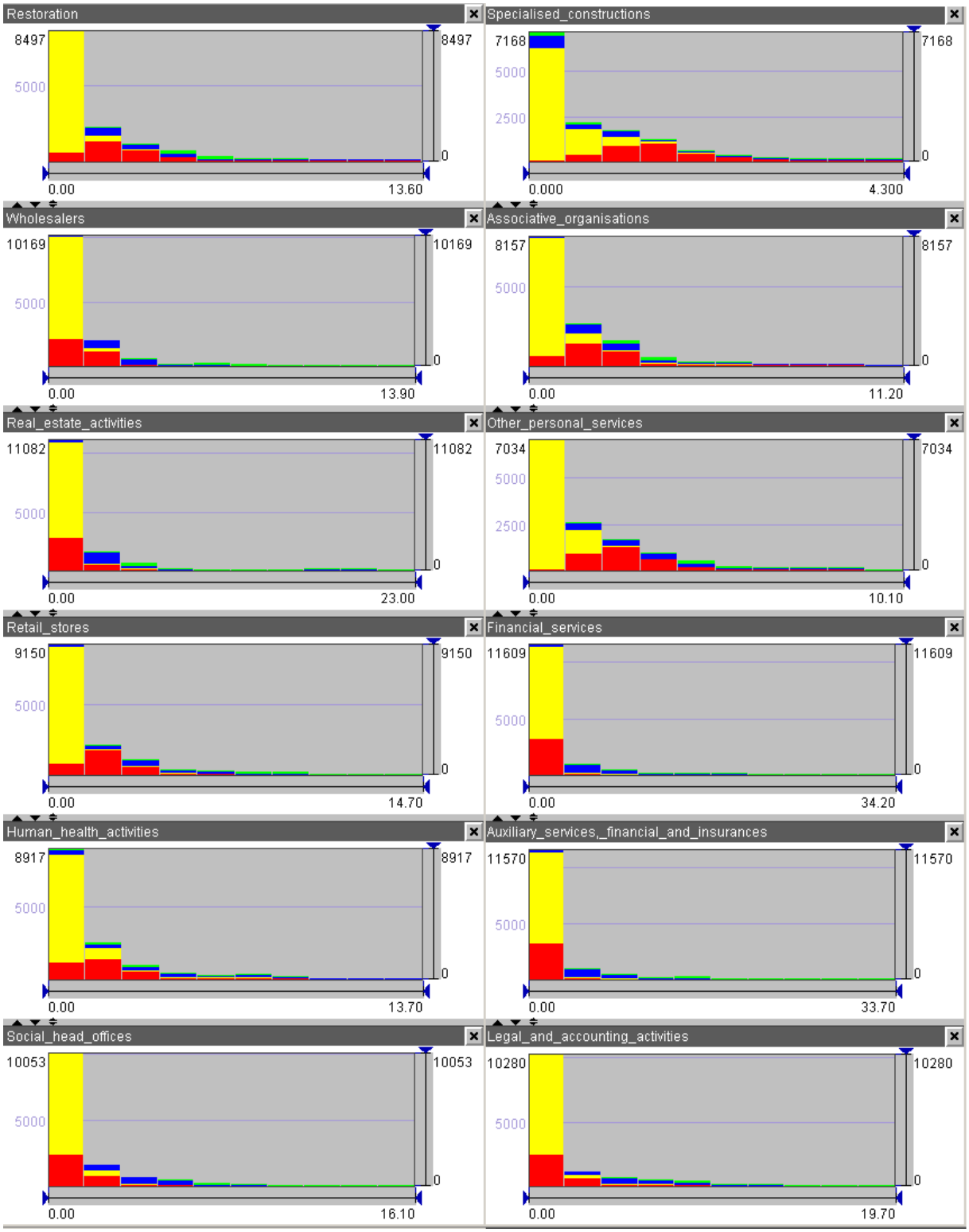

Figure 32: Cumulative histogram (10 bins) of the different classes generated with a Simple K Mean Clustering Method

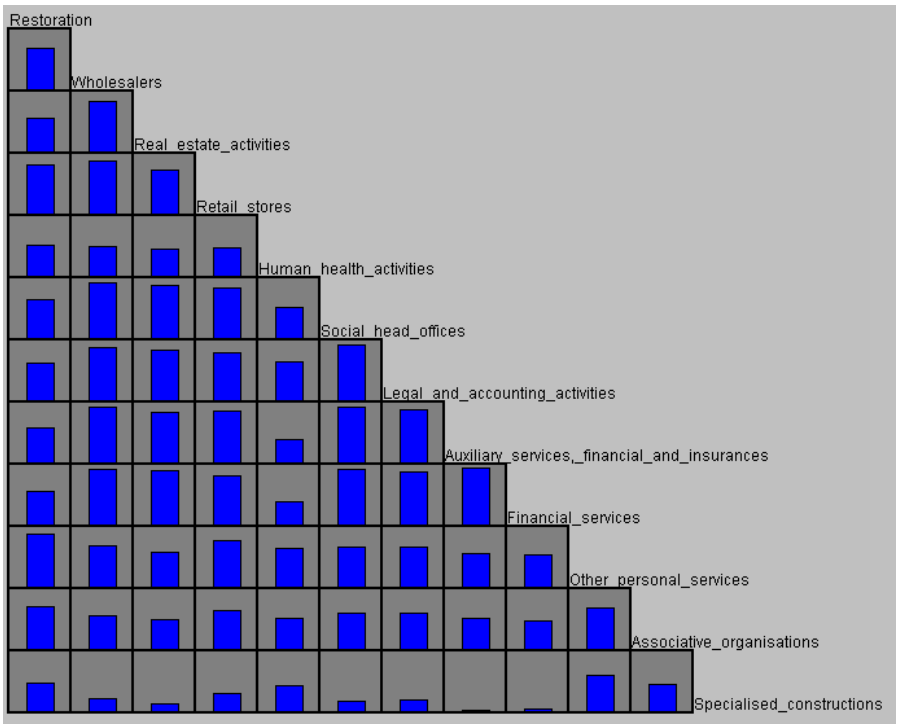

Figure 33: Correlations between densities of types of economical activities

7.4 Discussion and prospects

7.4.1 The "building" model

This chapter shows that by using the building footprints and by measuring distances exclusively on the network, it is possible to precisely characterize the density of economic activities in the direct neighborhood of a building. When generating a regular grid, one does not take into account the natural arrangement of the built area. As a result, points are randomly located in the city. As shown in the previous chapter, the absolute location of a grid of points is crucial in the calculation of the densities, in particular if one uses a coarse grid and/or NetKDE closest (segmentation of Euclidean space). The best way to remedy this problem, so far, was to choose a fine resolution grid and to use the "NetKDE visible" approach. The "building model" is therefore certainly the most accurate way to measure the density of economic activities in a city. Nevertheless, the assumption that the projection of a building centroid on its closest segment is the ideal starting point to compute the shortest-paths to all the spatial actors should be assess. As certain buildings possess complex shapes and the resulting centroids could sometimes be generated in unexpected locations. Moreover, some buildings are located between two roads or more, and it would therefore be possible to access either of these roads in reality. The ideal solution would be to identify the exact location of the entrances of the buildings and to perform a sensitivity analysis. One could also divide the construction into several smaller pieces and

calculate the resulting densities for each point. The variability between both models could then be computed.

The polygons generated on the basis of the SPT are first presented as a measure of the accessible surface. It has been shown that these shapes can be used to replace the traditional search area of a disk when measuring a network density or generating diversity indices. However, it would possible to generalize the use of these polygons to other indices whenever search areas are generated in a network space. Using a validation method, one would assess the advantage of using one approach instead of another.

7.4.2 Clusters of economic activities

It has also been demonstrated that by combining the density values of different types of economic activities, it is possible to determine categories where different kinds of spatial actors are grouped together. By looking at the correlations on Figure 33, one has an idea of which activities are more likely to be found together. It is, therefore, a good way to determine the relative attraction or repulsion of economical activities. The classification has been performed on a limited number of activities and at a relatively low level in the NOGA classification, but one could use a more detailed level of classification and perform a complete analysis of the situation.

The goal of this analysis is an attempt to determine where and how economic activities are grouped together. However, if the aim was to evaluate the relative importance of a sector of activity over another, one should weight the densities with the number of employees. The resulting values would then represent the economic importance of a type of activity in the city. This feature would reduce the impact of the class of “retail stores” and other parts of the city would be highlighted.

7.4.3 Hedonic prices

As introduced earlier in this chapter, one of the major advantages of calculating densities on the buildings is the possibility to set this information against other building attributes. Indeed, one could study the impact of the neighborhood attributes on the buildings’ residents or on the rent prices. A few studies have been aimed at integrating neighborhood factors in hedonic price modeling (Des Rosier, Thériault, & Villeneuve, 2000)(Thériault, Des Rosiers, Villeneuve, & Kestens, 2001). In fact, it has been pointed out that a substantial portion of price variability remains unexplained and that spatial statistics method could help integrating additional neighborhood factors to explain market variability.

7.4.4 Towards a new walkability index

Researchers recently pointed out the fact that features of the built environment may have influences adult's physical activity and probability of obesity (Rundle, et al., 2009). Indeed, neighborhood attributes may be potentially important for future environmental and policy initiatives designed to increase physical activity. The field of investigation to determine the influence of neighborhood factors on people health is extended. For instance, it has been demonstrated in New York that a higher local density of BMI-healthy food outlets accounted for a small but statistically significant lower prevalence of overweight and obesity (Rundle, et al., 2009). The presence of retail facilities, grocery stores and restaurants are also strongly associated with walking sufficiently to meet recommendations for health (Moudon, et al., 2006). In this context, GIS can be used as an objective measure to help better understand the relationships between physical environment attributes and physical activity behaviors (Leslie, Coffee, Lawrence, Owen, Bauman, & Graeme, 2007). It has also been pointed out that the spatial unit commonly used would need to be readjusted significantly downward to better capture the neighborhood walkability (Moudon, et al., 2006). The tool developed during this master could help meeting these requirements and characterizing relevant neighborhood attributes. A "walkability" index has been developed to try to evaluate the induced impact of physical elements of local environments on walking behaviors. The method presented by Leslie et al. (2007) to compute a walkability index is described below, together with insights on how to adapt this technique to the "building model". The factors taken into account to build this index are the street connectivity, the dwelling density, the land-use diversity and the net retail area. This study is characterized by the use of the quadrat method in order to evaluate the walkability indexes. Therefore, the street network is not directly taken into account. This is of course one model among many others, but it shows how the tool developed could be easily integrated in this type of studies.

The **dwelling density** is defined as the number of dwellings per residential land area in a square of one kilometer. Scores from 1 to 10 are generated based on the deciles of the resulting densities. Using the accessible areas represented by the polygons and the zoning of the city, one can easily derive the density of dwellings per building neighborhood. High density neighborhoods are known to include mixed-use development which is likely to improve the access to a variety of complementary activities and thus, promote physical activity.

The **street connectedness** is derived by measuring the density of street intersections which degree (number of associated segments) is greater than or equal to 3 (Leslie, Coffee, Lawrence, Owen, Bauman, & Graeme, 2007). Nevertheless, this measure is calculated based on the number of intersections per square kilometer, and, therefore the network morphology is not directly taken into account. The resulting densities are classified into deciles and a standardized score from 1 to 10 is generated. The connectedness index presented earlier could easily be adapted to generate the same kind of scoring per building. High network connectivity positively influences the walking behaviors by providing a greater variety of potential routes and easier access to major roads where public transportation is available.

The **land-use diversity** is computed by measuring the entropy within a square of one kilometer of five classes of land-use types (residential, commercial, industrial, recreation and other). Once again, the resulting values are classified in ten classes to generate scores ranging from 1 to 10. The entropy measure has also been developed during this work and could be easily generated by selecting the appropriate number of classes. Diversity is important because diverse retail opportunities tend to make more frequent short shopping trips by walking (Leslie, Coffee, Lawrence, Owen, Bauman, & Graeme, 2007).

Finally, the **net retail area index** is computed by performing the ratio of the gross retail area divided by the total retail parcel area. This measure captures the degree to which retail is located near the roadway edge (Leslie, Coffee, Lawrence, Owen, Bauman, & Graeme, 2007). If this parameter has a low value, this would indicate that a significant part of the store is devoted to parking lots. On the other hand, values closer to 1 indicate that less space is devoted to cars. The scoring procedure introduced previously is then applied to the net retail area ratios. Rather than using a quadrat count, one could compute NetKDE on the class "retail stores" and weight the activities according to the ratio described above.

The resulting walkability index is then constructed by summing the scores obtained for the previous indexes and therefore ranges from 4 to 40.

8 Spatio-temporal evolution of IED Explosions in Baghdad

8.1 Introduction

The analysis of terrorist attacks patterns has been a topic of growing interest in the scientific community. Recent publications claim that terrorist attacks should be considered as a type of crime because the decision-making process of terrorists and criminals are considered quite similar (Townsley, Johnson, & Ratcliffe, 2008) and, therefore, terrorist attacks follow some of the same basic principles which can be used in the description of crime.

So far, most crime patterns analyses have focused on analyzing predefined time intervals (Johnson, et al., 2007). Tobler's first law of geography is "everything is related to everything else, but near things are more related than distant things." When applying this statement to criminal events or terrorist attacks over time, it is logical to assume that close events are more similar than distant ones. When studying fixed time intervals, one assumes that all events have the same weights over time. Nakaya and Yano (2007) highlighted the fact that this kind of approach could mask important aspects of crime clusters, particularly in differentiating stable and fluid hotspot (Johnson, Lab, & Bowers, 2008), as each type of hotspot requires a different kind of intervention when being detected.

Twonsely et al. (2008) have recently analysed the spatio-temporal features of improvised explosive devices (IED) in Iraq. They studied attacks taking place in 2004 over a 3 months period. According to their study, events showed clustering properties in space and time, suggesting that identifying these patterns might help to predict the risk of the subsequent attacks and, thus, provide more effective methods for countering terrorist violence. These studies identified the region at highest risk for a future attack to be at a distance of up to 1 km for period of up to 2 days following the original event. To determine this information for a set of data, each observation is compared with every other, and the space and time diffrences are computed. Then, to assess the importance of spatial (and/or temporal) randomness of attacks, a Monte Carlo analysis is performed by shifting the dates amongst the set of events, which allows for the creation of new realizations. Townsley et al. (2008) also sugget to map sequentially incidents in time, to identify and understand attack patterns and diffrentiate stable hotsptots from fluid hotspots.

8.2 Methodology

The data used to perform the spatio-temporal simulations are the IED explosions listed in Bagdad from the 1st of January 2004 to the 30th of March 2010 and have been imported from WikiLeaks. During this time period, there were 2'652 events in an area of about 1'030 km^2.

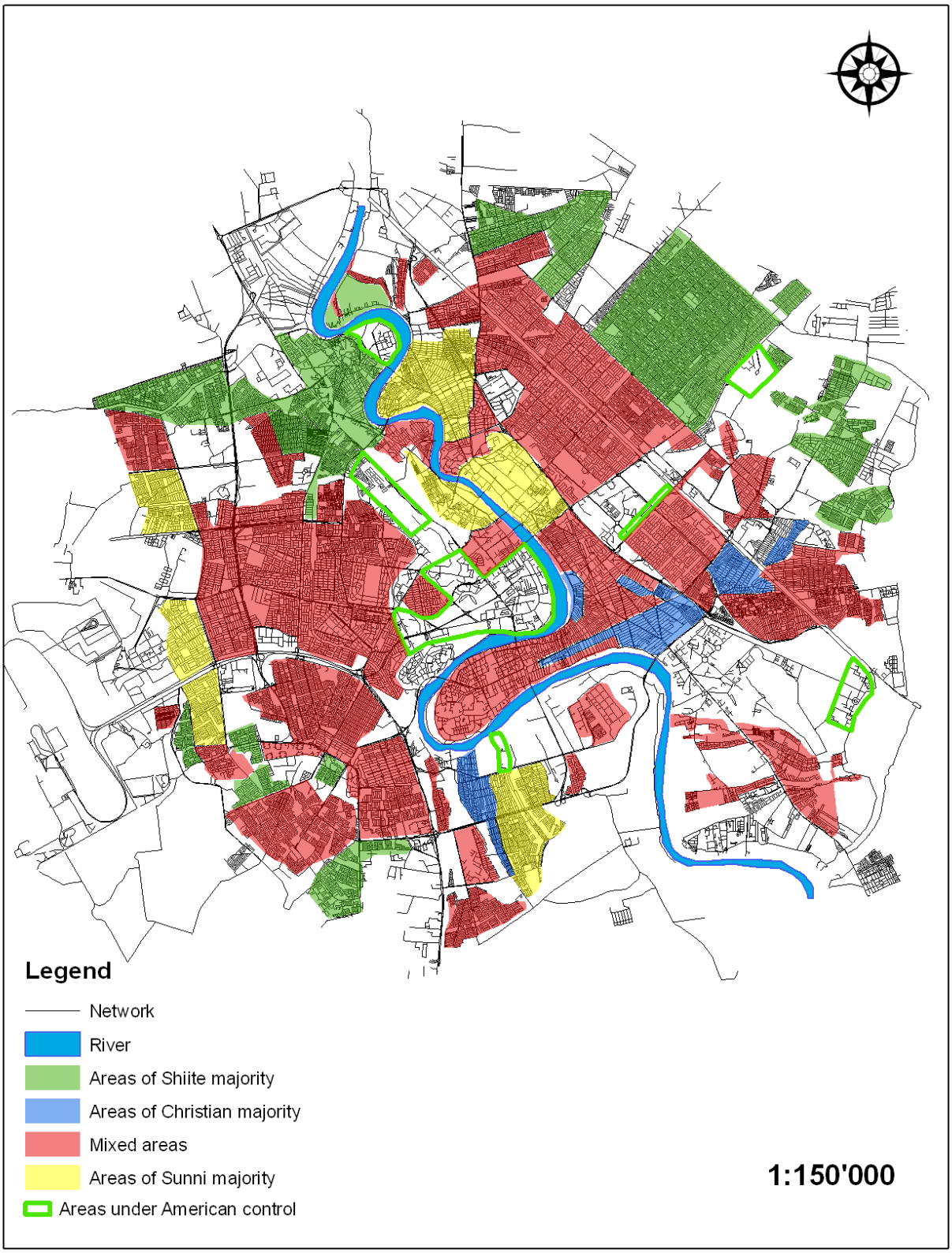

Figure 34: Distribution of religious groups

The distribution of religious groups has been approximated using maps found on the internet (Healing Iraq) and is presented on Figure 34. Based on these distributions, the numbers of events, as well as the density of events within each religious group have been calculated. The results are presented in Table 7. One can see that the highest density of events is found in the Christian group, while the lowest density is found in the Shiite Muslim group.

	American control	Sunni majority	Christian majority	Shiite majority	Mixed areas	Undefined
Number of Events	30	354	140	243	1085	800
Relative importance	1.13%	13.35%	5.28%	9.16%	40.91%	30.17%
Surface [km^2]	18.40	31.01	10.58	67.18	156.66	
Density [event/km^2]	0.128	0.171	0.369	0.083	0.052	-

Table 7: Number of attacks per zone

The aim of this chapter is to attempt to predict, as accurately as possible, where the new attacks will be localized the future. As previously introduced, it appears that the events are correlated in space and time. Assuming that there are plenty of potential targets in Baghdad, the distribution of points in space is not constrained. Therefore, the selection of an appropriate bandwidth is definitely the main subject of concern. One has to choose not only a radius of influence in space, but also in time.

In order to choose the most appropriate bandwidth, one must consider the attacks taking place only in the past. To do so, all the density surfaces are generated, using a predefined time step for all the possible combinations of bandwidth in time and space. Subsequently, the location of all new events taking place during the next time step, are stored in memory. For each new attack, the associated density value of the previous step is referenced.

This study is only focused on the IED explosions of three months, due to the overwhelming amount of data available. The distribution of explosions per month is presented on Figure 35. It can be seen that in 2006 and 2007 the number of attacks is significantly higher than the rest of the time period presented.

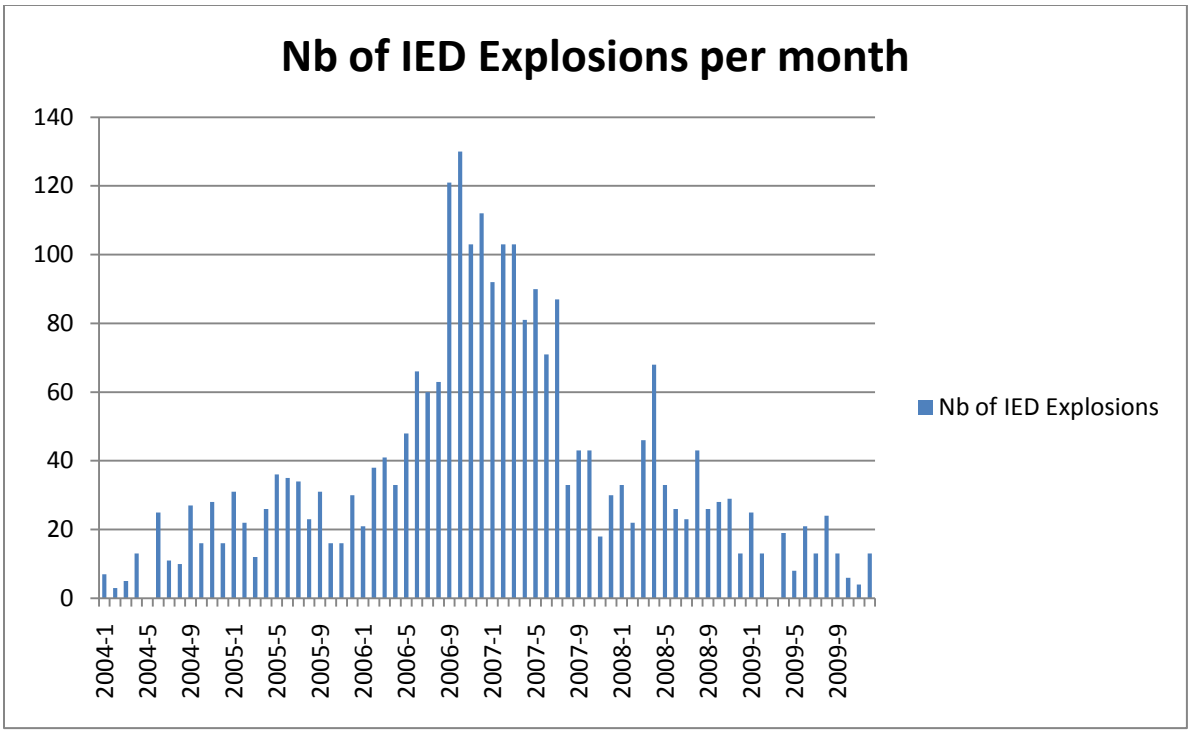

Figure 35: Cumulative histogram of the number of events per month

The maximum number of attacks takes place during the month of October 2006. As a result, it has been decided to select the explosions from the 1st of September to the 30th of November 2006. During this period, 30 different states for each bandwidth choice were generated with a time step of 3 days. The spatial resolution of the grid is 50 meters, and there are a total of 413'224 centroids. The kernel function used for the simulations is the Epanechnikov function defined in 3 dimensions.

An empirical study was carried out in order to select the bandwidth in space and time. For each new explosion, one looks at the associated density values and the surface occupied. First, the ranges of all non-zero density values are sorted in ascending order. Then, the value of the grid point associated to the event being examined is extracted, and its location in the range of values is determined. The densities equal to zero are assigned of value of 0%. Secondly, one looks at the entire surface occupied by the density values. As it will be shown later, the smoothing parameter choice is motivated by the aggregations of these two indicators for each bandwidth test.

In order to compute an integrated spatio-temporal function, a common unit is needed. Therefore, time has been normalized by the spatial bandwidth. Thus, for a spatial bandwidth of 1'000 meters and a temporal bandwidth of 100 days the correction factor value is equal to 1 day per 10 meters, which means that a time interval of 1 day is equivalent to a distance of 10 meters in the kernel function (Figure 36).

The correction factor can be expressed with the following equation

$$\alpha = \frac{h_t}{h_s} \quad (11)$$

where h_t is the temporal bandwidth, h_s the spatial bandwidth and α the correction factor. The spatial and temporal distances interact with one another and the value of one dimension determines the maximal value of the other. This relationship can be expressed with the following equation

$$d_{tMax} = \alpha\sqrt{h_s^2 - d_s^2} \quad (12)$$

where d_{tMax} is the maximal temporal interval taken into account for a spatial interval d_s. This equation has been computed for our case study and is presented on Figure 36.

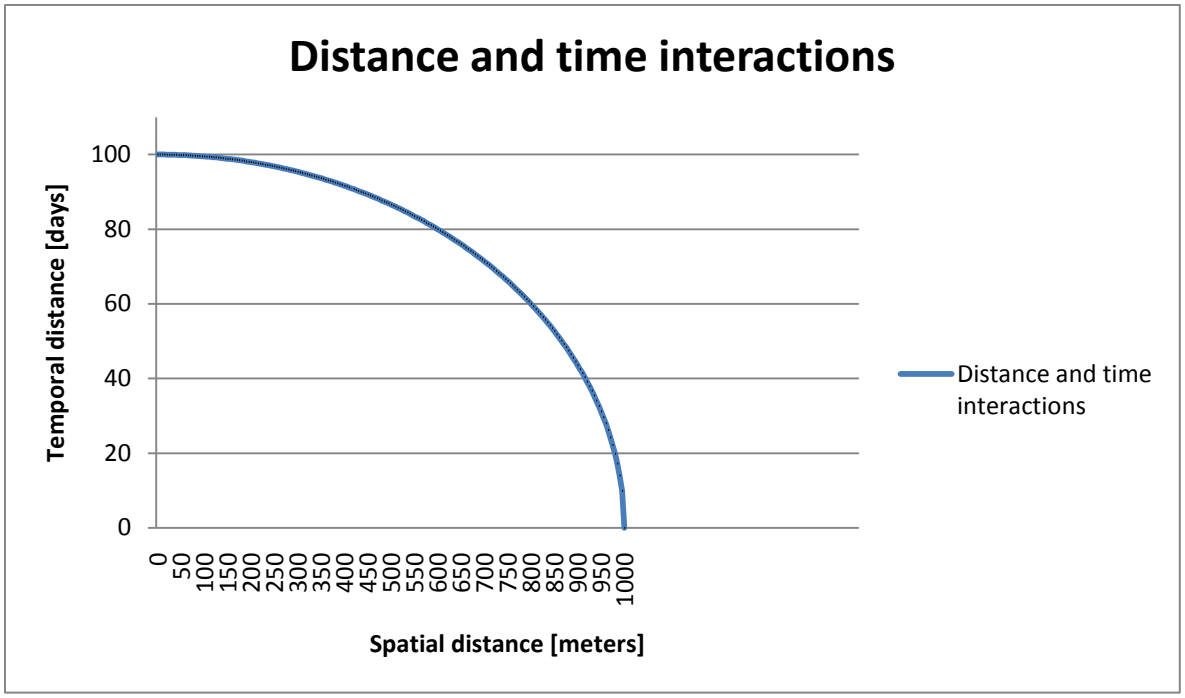

Figure 36: Spatio temporal interactions

The animations were generated in ArcMap using the module "animation", and a regular time step of 0.3 seconds was set between each image.

8.3 Results

The empirical approach used to select a bandwidth in space and time will first be introduced and then, an animation is proposed as a visualization approach of the results.

The values extracted from the predictive model described above are presented on Figure 37 and Figure 38. When looking at the results of the mean value for the events falling in the range of density values of the preceding state, one can see that when the spatial bandwidth increases, the number of attacks detected increases as well. This fact is not very surprising, considering that when one increases the bandwidth, one gives higher weights to closer events. But when considering the surface occupied, one can appreciate the opposite trend.
Therefore, the choice of the bandwidth is a tradeoff between an excessively high bandwidth which would better explain most of the data points, however, imprecisely, and a very small bandwidth which would only explain a few points but with high accuracy. Therefore, the decision of an appropriate window width is based on the search of the optimal compromise rather than the best solution.
On Figure 37, the temporal bandwidth becomes increasingly important along the spatial bandwidth gradient. Moreover, increasing the temporal bandwidth will always result in an increase of the mean values of the predictions. On Figure 38, higher bandwidths in time and space correlate with an increased percentage of surfaces covered by density values.
By combining these two graphs, one can estimate how accurate a bandwidth selection (in time and space) will be and what percentage of the study area will be occupied by density values. Nevertheless, the impact of changing the spatial or temporal bandwidth is hard to interpret on these graphs. Therefore, rather than looking at absolute values, one can compute the change of mean value and covered surface between two different bandwidth choices.

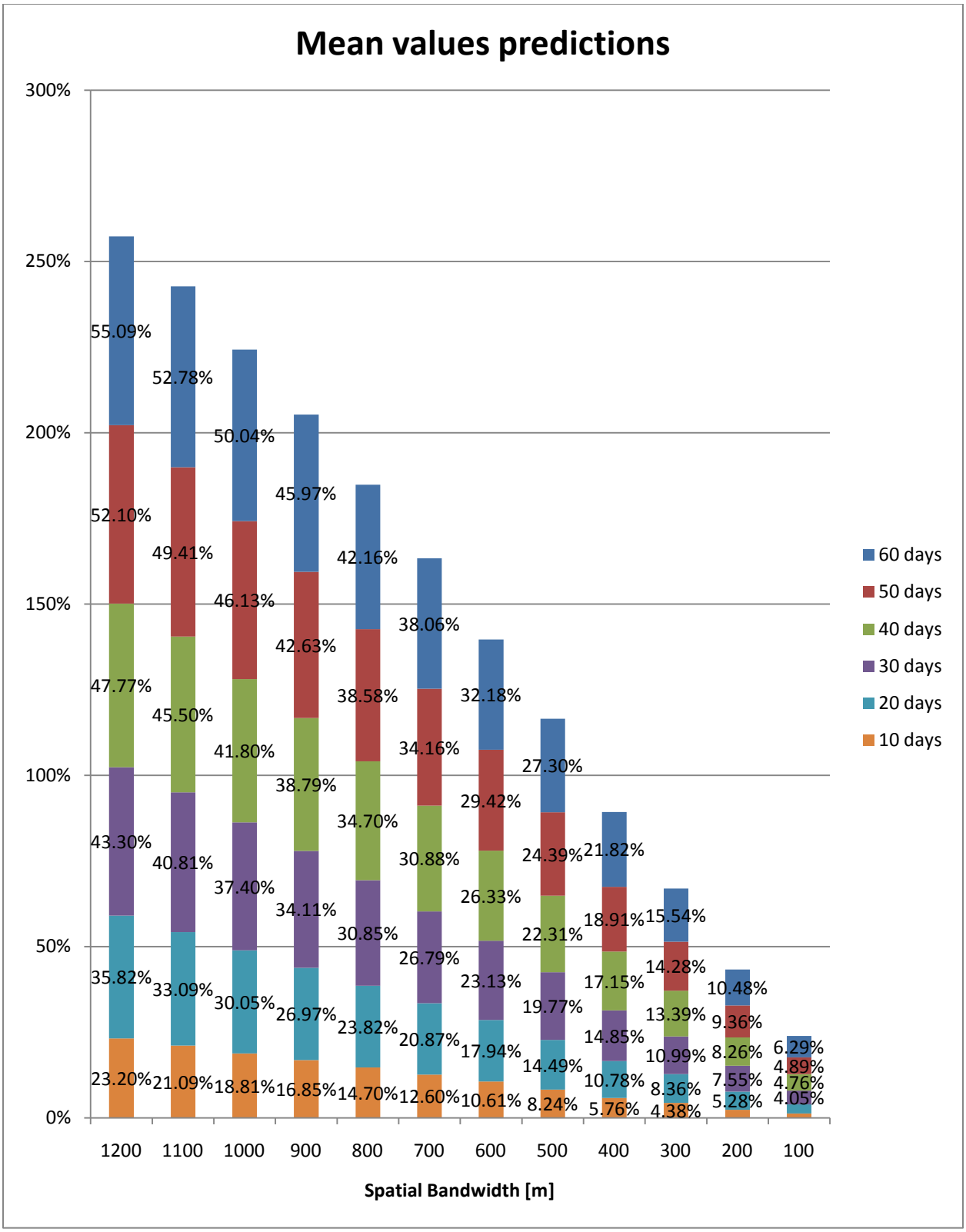

Figure 37: Cumulative histogram of the mean value of the total number of events falling in the range of the density values of the preceding state

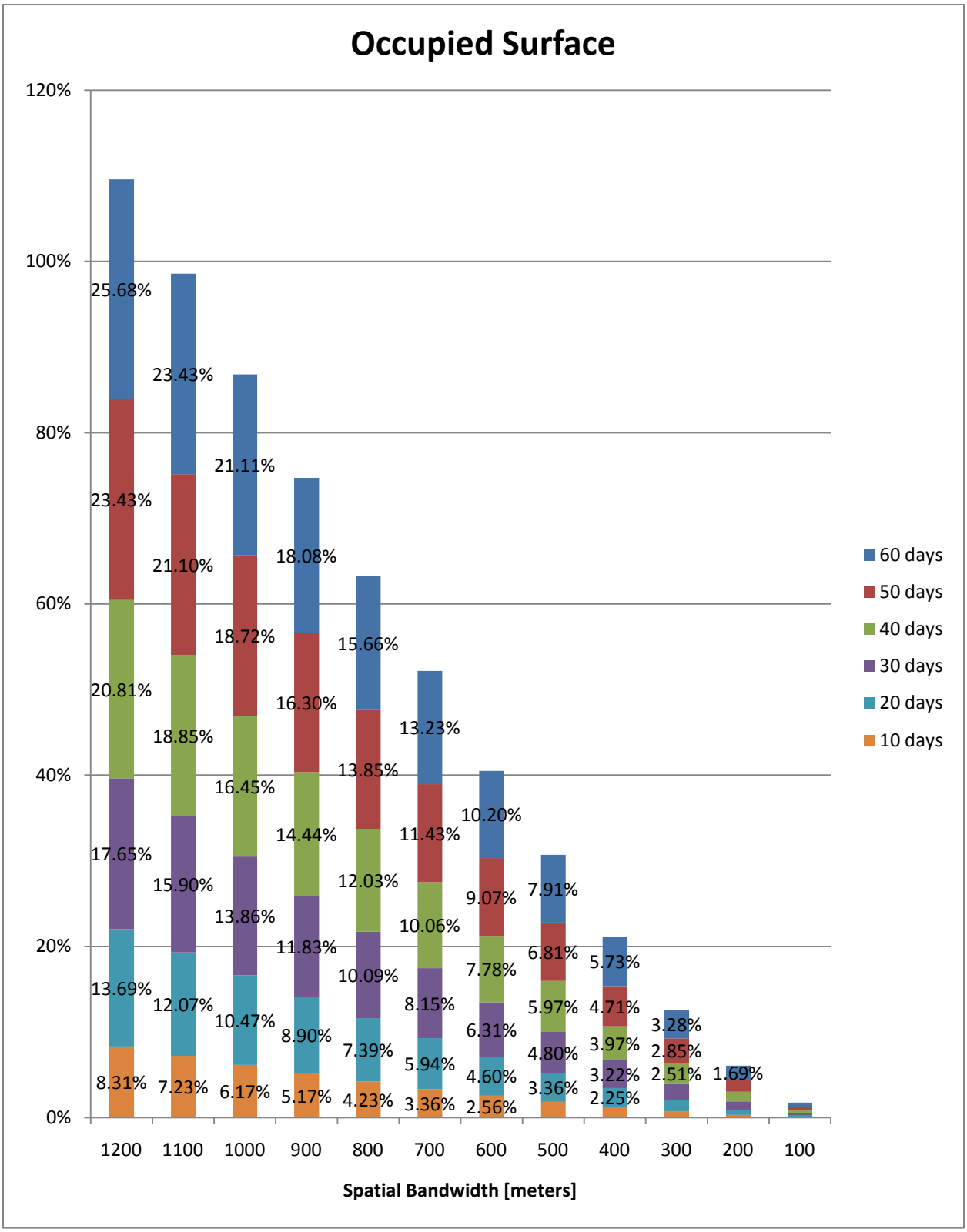

Figure 38: Cumulative histogram along the spatial bandwidth gradient according to the occupied surface [%]

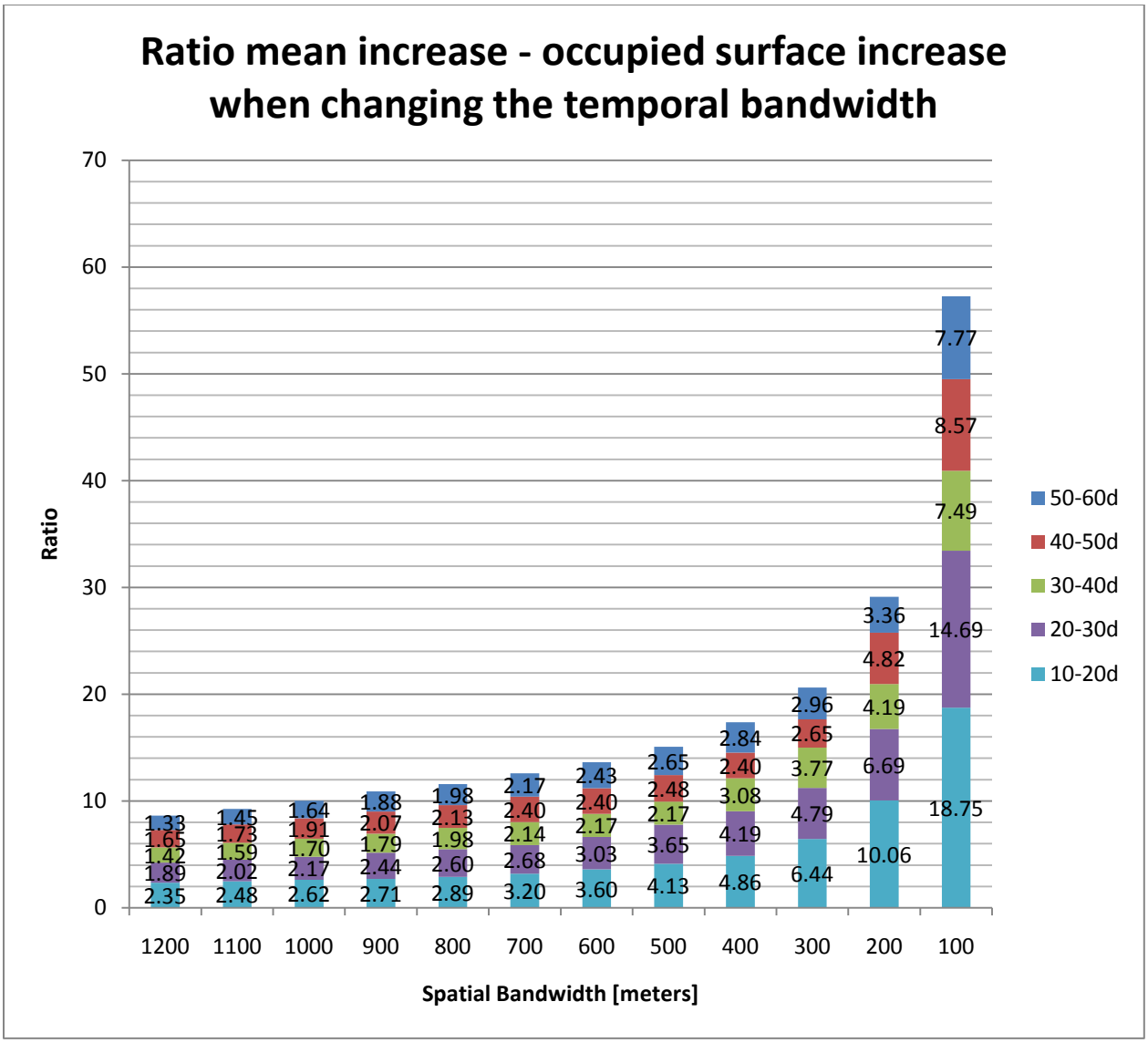

Figure 39: Ratio mean increase - occupied surface when changing the temporal bandwidth

On Figure 39 above, one can appreciate the ratio per spatial bandwidth between the increase of mean value of the evens falling in the range of values of the preceding state and the augmentation of occupied surface when increasing the temporal bandwidth. Therefore, an increase of the mean value is always normalized with the associated increase of occupied area. The highest values of this ratio are generated when the increase of occupied surface is minimized and the increase of the mean value of the predictions is maximized. For a smoothing parameter of 100 meters, the influence of the temporal window width is very high. Therefore, one can see that as the radius of influence of a previous event decreases, the influence of the time interval becomes more important. Indeed, between 100 and 500 meters, the overall influence of the temporal window width is at least divided by 4. For a fixed spatial bandwidth, one can see that the ratios don't increase regularly over time. This is an important feature for determining

temporal clusters. Indeed, the discontinuities between time intervals are certainly one of the most interesting information that one can get about this graph. The influence of raising the bandwidth from 10 to 20 days is always higher than any subsequent increases. This fact is also verified between 20 and 30 days. Nevertheless, between 30 and 50 days, the ratios do not always decrease along the temporal gradient (e.g. vertically). The ratios between 40 and 50 days are usually higher than between 30 and 40 days expect for bandwidths of 300 and 400 meters. As a result of the outcome of this analysis, it is recommended to select a temporal bandwidth of at least 30 days for spatial bandwidths smaller or equal to 500 meters. Using a window width of 50 days might also be interesting in order to highlight more general trends for higher spatial bandwidths.

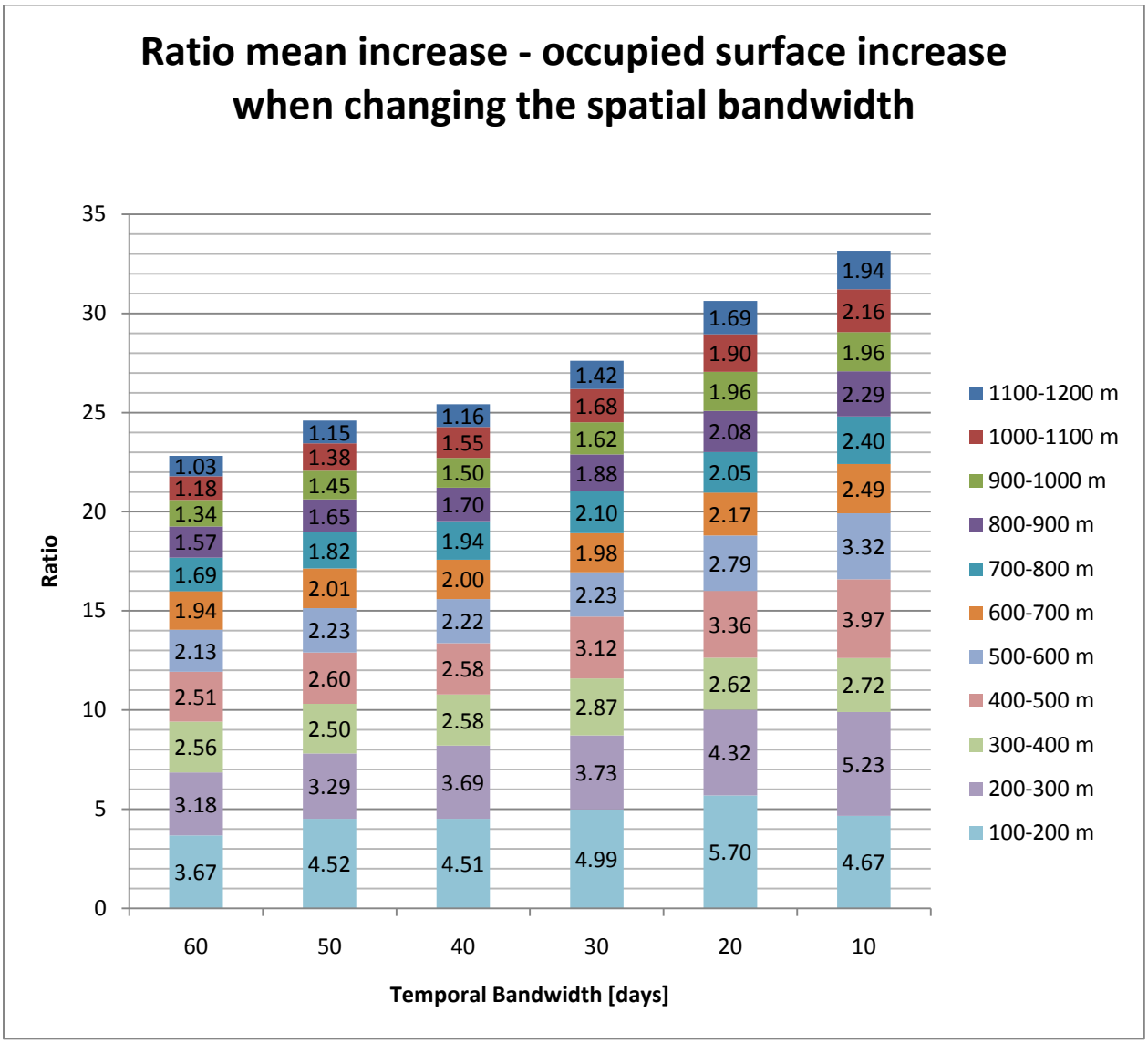

Figure 40: Ratio mean increase - occupied surface when changing the spatial bandwidth

If now one looks at the evolution of this ratio when one fixes the time interval and increases the spatial bandwidth, one should be able to locate the clusters approximately. The results are presented on Figure 40 hereunder. The amount of information one is able to get when increasing the radius of influence in space from 100 to 300 meters is relatively high for all the temporal bandwidths. While the main observable trend indicates that increasing the spatial bandwidth results in a decrease between the differences of the two states, it is interesting to point out that when one increases the radius from 300 to 400 meters, the relative information that one gets about the explosions is lower than the information between 400 and 500 meters. The same phenomena can be observed between 700 and 800 meters for a time interval of 30 days. This analysis indicates that one would miss important features of the explosions distribution, if one does not take into account events in a bandwidth of at least 500 meters. A radius increase from 100 to 500 meters represents at least half of the total augmentation of the ratio.

As a result of this analysis, it has been chosen to select a temporal bandwidth of 30 days and a spatial bandwidth of 500 meters. This choice represents the minimal value that one should consider to perform a spatio-temporal KDE analysis. On average only 4.8% of the total study area is covered, which represents a total surface of about 49 km^2. This area is already quite extended if one assumes that eventual measures should be taken primarily in these zones to counteract terrorist attacks. The scale at which the events are considered is therefore intimately related to the area the Iraqi and the Coalition Forces are able to control. On average the events fall in the 2^{nd} decile of the density values of the preceding states and 20 are above the 8^{th} decile. The distribution of the events prediction over time is presented on Figure 41.

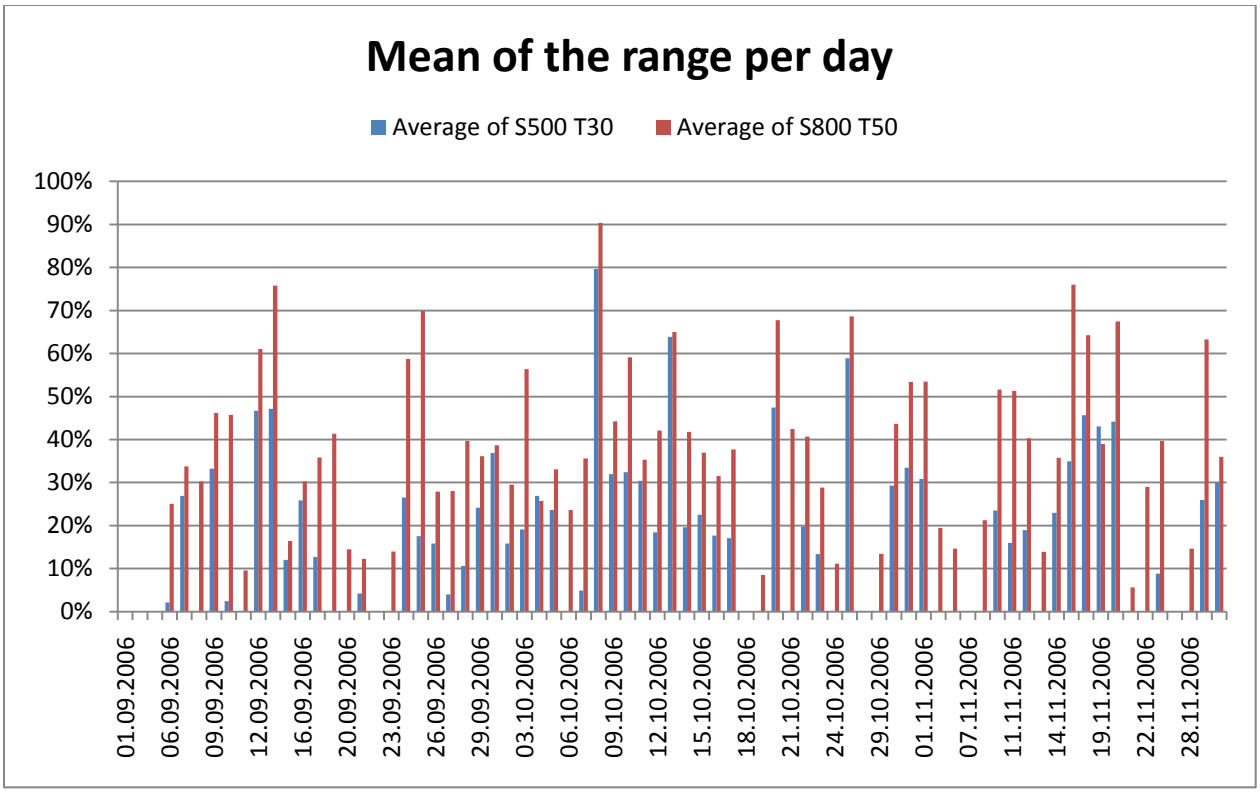

Figure 41: Mean value of the predictions per day (bandwidth of 500 m and 30 days in blue, 800 m and 50 days in red)

Therefore, one can see that the quality of the events prediction is not constant along the time line. While a bandwidth of 500 meters and 30 days the events occurring between the 7th and the 17th of October 2006 seem to possess relatively high prediction values, the events occurring between the 2nd and the 9th of November 2006 are all out of the prediction zones. With a bandwidth of 800 meters and 50 days, the events start to enter in the prediction zone at the beginning of November. By using such a bandwidth, the surface covered is tripled and the average value of the predictions doubled. When the bandwidth is modified, different clusters at different scales can be identified. Figure 41 can be used with the animation to assess where and when the events are predicted correctly.

When visualizing the animation, the number of attacks increases significantly during the month of September, and it seems that starting from the north from the area 1 on Figure 42, the attacks slowly move towards the south near the zone 2 on Figure 43 (following the arrow). When looking at the quality of the predictions (Figure 41) for this month, they are very good for the 9th and the 10th of September. The attacks took place nearly exactly at the same location than a couple of days before. Between the 17th and 28th the prediction values are quite bad. In fact, this period matches with the displacement of the attacks towards the South.

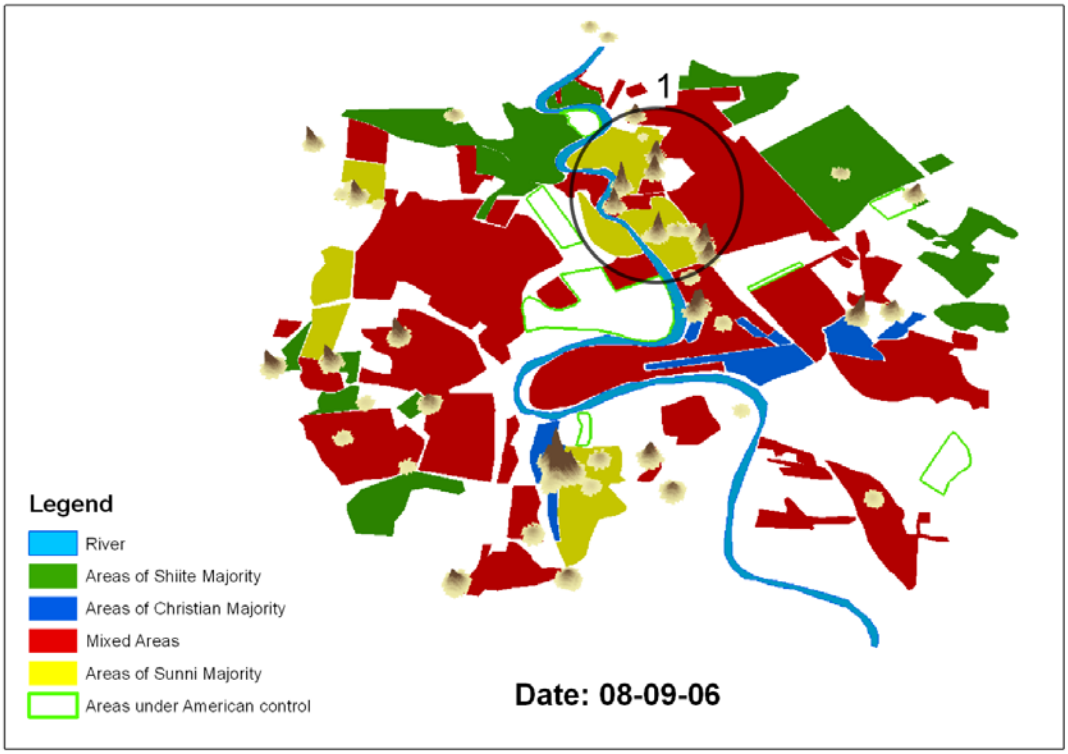

Figure 42: IED explosions in Baghdad (08-09-06), 500 m 30 days

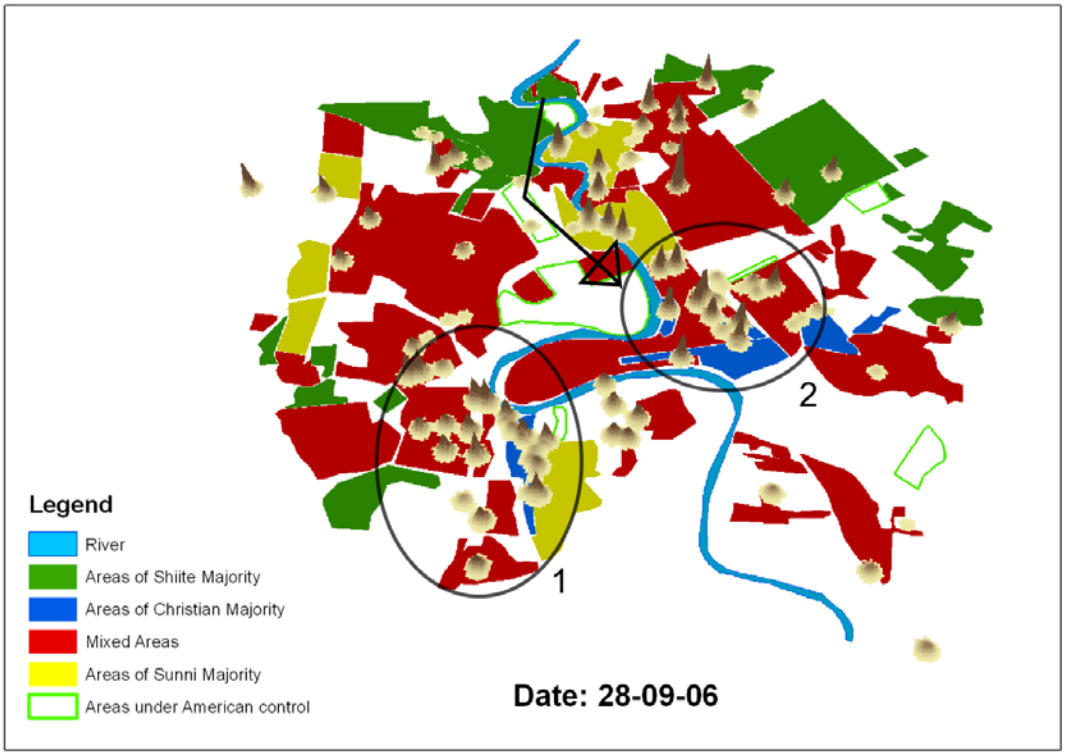

Figure 43: IED explosions in Baghdad (28-09-06), 500 m 30 days

Until at least the 15th of October, the events are geographically close to one another. This trend can be seen in Figure 41. The events possess a higher mean value for this time period. Between the middle of the month of October, till the end of the same month, the attacks move slowly in the direction of the north on both sides on the Tigris. This phenomenon is represented in Figure 44 below. The areas 1 and 2 in Figure 45 are the hotspots at the end of October.

Therefore, the explosions are mainly located on two large stripes on each side of the river. It seems that a periodic displacement of the events occurs over time. The areas of Christian majority and the areas of Sunni majority seem in general more prone to the attacks for the time interval studied so far. The fact that the areas under American Control are never under attack during these three months, shows how difficult it is to reach these zones for the terrorists.

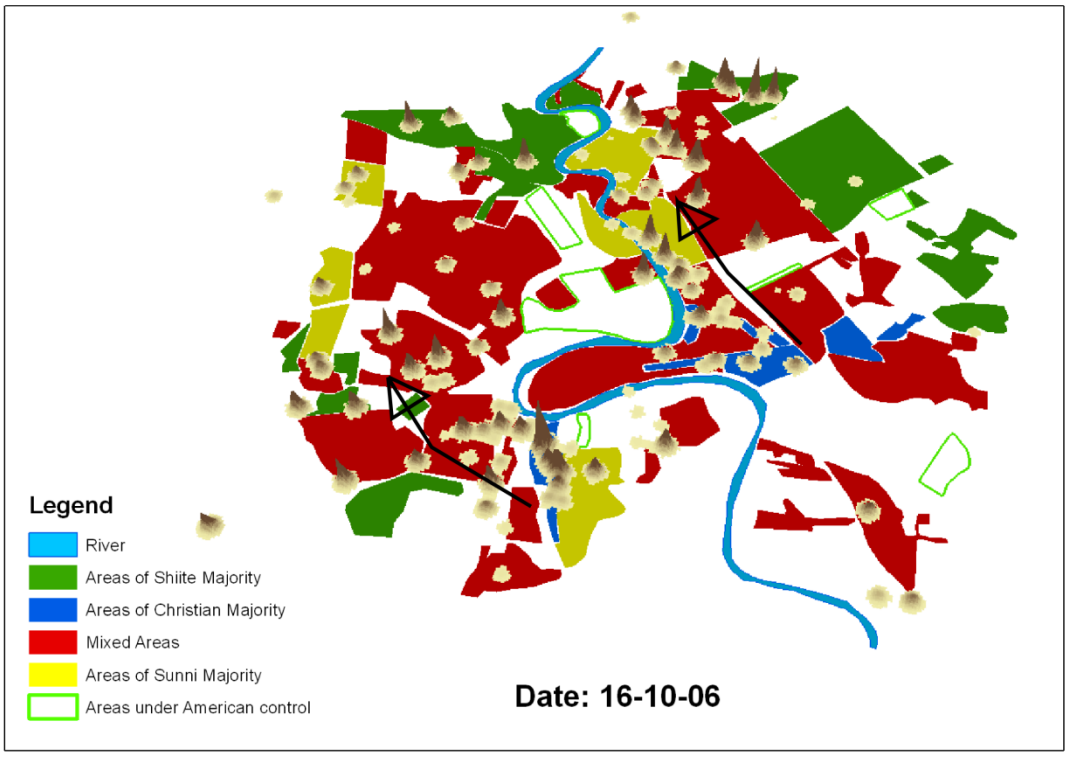

Figure 44: IED explosions in Baghdad (16-10-06), 500 m 30 days

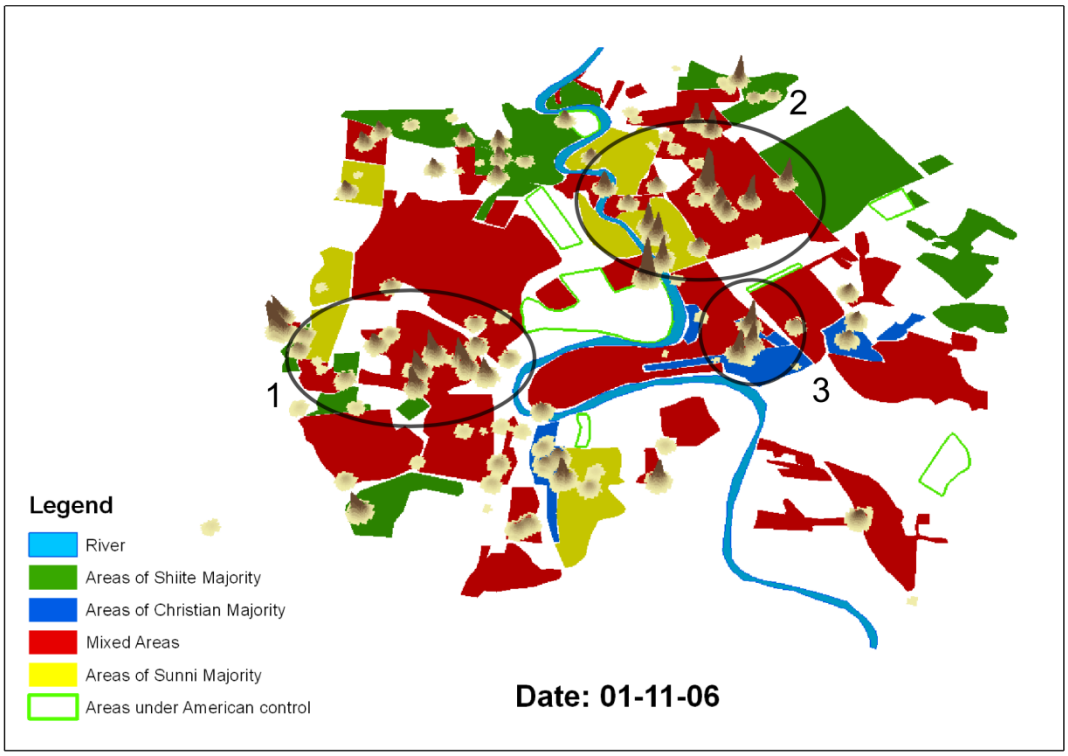

Figure 45: IED explosions in Baghdad (01-11-06), 500 m 30 days

8.4 Discussion and prospects

8.4.1 Bandwidth selection

In this chapter, it has been demonstrated that the events chronology can be taken into account by integrating the temporal dimension in KDE. The empirical approach used to determine the bandwidth possesses the advantage of showing explicitly the impact of changing the bandwidth on the resulting density surfaces. It has also been demonstrated that the rate of change between different bandwidths is not identical. This indicates that there may be a correlation between the events taking place in Baghdad. By studying the distribution of events for different bandwidth in time and space, it could be possible to identify clusters at different scales. The same kind of analysis could be carried out with different kernel density parameters. One could change the kernel function for instance. It would be quite easy to compare two different simulations with the same bandwidth parameters, and assess the predictive pertinence of each kernel function. One could also evaluate the differences between the separated (e.g. density values distributed over a cylinder) and the integrated spatio-temporal kernel density estimation. Nevertheless, this method still requires improvements in order to use it as an effective prediction tool.

8.4.2 The animation

It has been demonstrated that by using an animation to represent the IED explosions in Baghdad through space and time, the main features of the attacks could be distinguished. The relative periodicity observed during the months of September and October would have been difficult to notice without any dynamic visual support. Therefore, animations can be helpful at characterizing fluid hotspots. Other visualization methods should be assessed in order to define the optimal technique, or set of techniques to explore spatio-temporal data.

8.4.3 Network integration

When visualizing the distribution of the IED explosions in Baghdad, similarly to crime events, they are mostly distributed along the road network. Indeed, most of the attacks usually take place in the street where crowds are gathered like in market areas. 1'830 of the 2'652 identified events involve the death of at least one civilian. Concerning the other targets of such attacks, Iraqi and Coalition forces are more vulnerable when moving along the road. In fact, nearly no explosion occurred in the areas controlled by the American forces. Therefore, it is possible to make the assumption that the locations of IED explosions are constrained by the network, similarly to the built area. Moreover, the city of Baghdad is crossed from the north to the south by the Tigris which acts as a physical barrier between both sides of the city. Twelve bridges are distributed along the river and enable the communication between the east and west side. As a conventional KDE fails at integrating these features, it could be interesting to compute NetKDE over a 3 dimensional space.

8.4.4 Other possible applications

The present application of 3D KDE has been presented as a prediction tool to assess the locations of future terrorist attacks. This method might have potential applications for similar fields of investigation (Nakaya & Yano, 2010). Nevertheless, this approach could certainly be used in a number of other fields, and represent various kinds of information. For instance, any displacements localized in space and time by point events could be analyzed by this mean. Therefore, as it has been performed to visualize monthly vessels trajectories (Demsar & Virrantaus, 2010), one could use this method to study displacements of human or animal populations at different time scales. As GPS tracking techniques become increasingly popular (Lachance-Bernard, Produit, Tominc, Matej, & Golicnik Marusic, 2011), the amount of data available is likely to rise rapidly. As a result, it will become necessary to find appropriate techniques to treat and extract information from this gigantic amount of data. Even if this technique is still in his infancy, it seems a serious lead for managing such kind of data.

9 Conclusion

The new implementation of the tool enables easier exploration of the data. The user does not always know what kind of information he is looking for and, therefore, he might want to run several simulations on the same dataset and assess the impact of changing one or several parameters. By storing the distances between the centroids and the events, it is possible to generate new results much quicker. Due to the increased number of potential input parameters and the higher level of intermediate computations, the program is designed to theoretically adapt to various kinds of problems. The interface enables any user, independently of his programming experience, to analyze spatial data constrained by a network. Nevertheless, it should be pointed out that the program is not designed to analyze high distances on the network. Distances up to 2 km can normally be used without any issue. However, above this threshold, the time needed to access the database may increase dramatically. The size of the generated database depends, of course, also on the number of centroids and events being studied.

It has been demonstrated that by using a slightly different algorithm ("NetKDE visible") than the original one to compute NetKDE, it is possible to generate smoother density surfaces. By changing the original hypothesis (minimization of the Euclidean distance) to compute density values in the area where the network is embedded, it has been demonstrated that NetKDE can be less sensitive to the location of the centroids by performing a visibility analysis. Nevertheless, the use of the visible segment approach is not recommended in all the situations. Linear networks, like rivers, do not take advantage of the characteristics of the new algorithm. Moreover, depending on the size of the dataset, this approach could be too time consuming.

A method combining the building footprints and shortest-paths measured on the network to analyze point processes has been presented. In an urban context, this approach seems to better integrate the morphology of the city and, therefore, would provide potentially more accurate results. In fact, in the future, thanks to the increasing computational power and research innovations, this kind of analysis might progressively become more widespread. As introduced earlier, this method could certainly be used in a number of other studies. A deeper understanding of the implications of using the centroids of the building footprints would be very useful to assess the quality and the pertinence of this model.

10 Bibliography

(n.d.). Retrieved 06 25, 2011, from Healing Iraq: http://healingiraq.blogspot.com/maps.htm

Biba, G., Des Rosiers, F., Theriault, M., & Villeneuve, P. (2006). Big Boxes versus Traditional Shopping Centers: Looking At Households' Shopping Trip Patterns. *Journal of Real Estate Literature* , 14 (2), 175-202.

Biba, G., Thériault, M., & Des Rosiers, F. (2007). Analyse des aires de marché du commerce de détail à Québec: une méthodologie combinant une enquête de mobilité et un système d'information géographique. *CyberGEO* .

Biba, G., Thériault, M., Villeneuve, P., & Des Rosiers, F. (2008). Aires de marché et choix des destinations de consommation popur les achats réalisés au cours de la semaine - Le cas de la région de Québec. *Le Géographe canadien* , 52 (1), 38-63.

Borruso, G. (2003). Network Density and the Delimitation of Urban Areas. *Transactions in GIS* , 7(2): 177-191.

Borruso, G. (2008). Network Density Estimation: A GIS Approach for Analysing Point Patterns in a Network Space. *Transactions in GIS* , 12(3): 377-402.

Borruso, G. (2005). Network Density Estimation: Analysis of Point Patterns over a Network. *ICCSA* , 126-132.

Brunsdon, C. (1995). Estimating probability surfaces for geographical point data: An adaptive kernel algorithm. *Computers and Geosciences* , 21 (7), 877-894.

Brunsdon, C. (2001). The comap: exploring spatial pattern via conditional distributions. *Computers, Environement and Urban Systems* , 25, 53-68.

Brunsdon, C., Corcoran, J., & Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems* , 31, 52-75.

Chhetri, P., Corcoran, J., & Stimson, R. (2009). Exploring the Spatio-Temporal Dynamics of Fire Incidence and the Influence of Soci-Economic Status: A Case Study from South East Queensland, Australia. *Journal of Spatial Science* , 54 (1), 79-91.

Demsar, U., & Virrantaus, K. (2010). Space-time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science* , 24 (10), 1527-1542.

Des Rosier, F., Thériault, M., & Villeneuve, P.-Y. (2000). Sorting out access and neighbourhood factors in hedonic price modelling. *Journal of Property Investment and Finance* , Vol. 18, No. 3, pp. 291-315.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik 1* , 269-271 .

Downs, J. A., & Horner, M. W. (2007). Effects of Point Pattern Shape on Home-Range Estimates. *The Journal of Wildlife Managemen* , 72(8): 1813-1818.

Downs, J., & Horner, M. (2007). Characterising Linear Point Patterns. *GISience Research UK Conference (GISRUK)* .

Downs, J., & Horner, M. (2007). Network-Based Kernel Density Estimation for Home Range Analysis. *GIScience Research UK Conference (GISRUK)* .

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Royal Geographical Society* , 21: 256–274.

Genre-Grandpierre, C., & Foltête, J.-C. (2003). Morphologie urbaine et mobilité en marche à pied. *Cybergeo : European Journal of Geography* .

Hsieh, T.-J., Chen, C.-K., & Ma, K.-L. (2010). Visualizing Field-Measured Seismic Data. *IEEE Pacific Visualisation Symposium* , 65-72.

Jansenberger, E. M., & Staufer-Steinnocher, P. (2004). Dual Kernel Density Estimation as a Method for Describing Spatio-Temporal Changes in the Upper Austrian Food Retailing Market. *7th AGILE Conference on Geographic Information Science* , 551-558.

Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J., Rengert, G., et al. (2007). Space-Time Patterns of Risk: A Cross National Assessment of Residential Burglary Victimization. *23 (3).*

Johnson, S. D., Lab, S. P., & Bowers, K. J. (2008). Stable and fluid hotspots of crime: Differentiation and identification. *34: 32-45*.

Kanungo, T., M. Mount, D., Silverman, R., Netanyahu, S. N., Wu, A. Y., & Piatko, C. (2000). The Analysis of a Simple k-Means Clustering Algorithm. *Computational Geometry* , 100-109.

Lachance-Bernard, N., Produit, T., Tominc, B., Matej, N., & Golicnik Marusic, B. (2011). Network based Kernel Density Estimation for Cycling Facilities Optimal Location Applied to Ljubljana. *ICCSA, Part II, LNCS 6783* , 136-150.

Leslie, E., Coffee, N., Lawrence, F., Owen, N., Bauman, A., & Graeme, H. (2007). Walkability of local communities: Using geographic inforamtion systems to objectively assess relevant environemental attributes. *Health and Plac* , 111-122.

Maki, N., & Okabe. (2005). A spatio-temporal analysis of aged members of a fitness club in a suburb. *Proceedings pf the Geographical Information Systems Association* , 29-34.

Moudon, A. V., Chanam, L., Cheadle, A. D., Garvin, C., Johnson, D., Schmid, T. L., et al. (2006). Operational Definitions of Walkable Neighborhood: Theoretical and Empirical Insights. *Journal of Physical Activity and Health* , 99-116.

Murgante, B., Borruso, G., & Lapucci, A. (2009). Geocomputation and Urban Planning. *SCI 176* , 1-17.

Nakaya, T., & Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. *14(3).*

Okabe, A., Satoh, T., & Sugihara, K. (2009). A kernel density estimation method for networks, its computanional method and GIS-based tool. *International Journal of Geographical Information Science* , 23 (1), 7-32.

Okabe, A., Satoh, T., Furutas, T., Suzuki, A., & Okano, K. (2008). Generalized network Voronoi diagrams: Concepts, computational methods, and applications. *International Journal of Geographical Information Science* , 22 (9), 965-994.

Produit, T., Lachance-Bernard, N., Strano, E., Porta, S., & Joost, S. (2010). A Network Based Kernel Density Estimator Applied to Barcelona Economic Activities. *ICCSA* , 32-45.

Rundle, A., M. Neckerman, K., Freeman, L., S.Lovasi, G., Purciel, M., Quinn, J., et al. (2009). Neighborhood Food Environement and Walkability Predict Obesity in New York. *Environmental Health Persepectives* , 442-447.

Silverman, B. (1986). Density Estimation For Statistics and Data Analysis. In *Published in Monographs on Statistics and Applied Probability.* London: Chapman and Hall.

Thériault, M., Des Rosiers, F., Villeneuve, P., & Kestens, Y. (2001). Modelling Interactions of Location with Specific Value of Housing Attributes. *Document de travail* .

Townsley, M., Johnson, S. D., & Ratcliffe, J. H. (2008). Space Time Dynamics of Insurgent Activity in Iraq. *Security Journal* , 21, (139-146).

Van Eck, J. R., & Koomen, E. (2007). Characterising urban concentration and land-use. *Springer* , DOI 10.1007/s00168-007-0141-7.

Wolff, M., & Asche, H. (2010). Towards 3D Tactical Intelligence Assessments for Crime Scene Analysis. *ICCSA* , 346-360.

Xie, Z., & Yan, J. (2008). Kernel Density Estimation of traffic accidents in a network space. *Computers, Envrionment and Urban Systems* , 32 (396-406).

11 Appendix

11.1 Dijkstra's shortest-path algorithm

The algorithm of the shortest-path has been developed by Edsger Dikjstra in 1959. In order to calculate the shortest path on a network, all the segments must have a starting node (FromNode) and an ending node (ToNode). As each segment is associated with two points having a unique Id numbers, from an initial node (starting node) on the network one can compute the shortest-path to a target node or until a maximum distance is reached (Dijkstra, 1959).

The algorithm would include the following steps:

1) From a starting node, set infinity distance to all other nodes and zero for the initial node.
2) Define the current node as the one that has the shortest recorded distance and consider the nodes connected to it. If the distance to this node is lower than the one recorded, overwrite the distance.
3) When all the neighboring nodes have been visited, mark the current node as visited (remove it from the list). The distance to this node is now definitive and it will not be visited again.
4) Dijkstra's algorithm ends when all the nodes have been visited or when a target node has been visited.

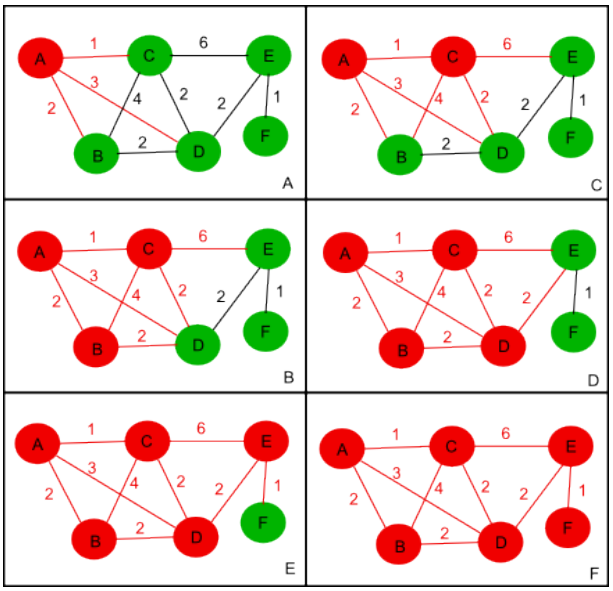

Figure 46: Dijkstra's shortest-path algorithm

Finally to implement the algorithm, one need:

1) A table that keeps track of the nodes visited
2) A table that keeps track of the effort to each node from the initial node
3) A loop that iterate until all nodes are visited
4) A function that checks the effort

Figure 46 is an example of Dijkstra's shortest-path algorithm. Table 8 shows the different steps of the algorithm. The numbers in red represent the shortest distance still in the list, and therefore the next node to be visited.

Node visited	Distance to A	Distance to B	Distance to C	Distance to D	Distance to E	Distance to F
Initialize	0	∞	∞	∞	∞	∞
A	0	2	1	3	∞	∞
C	0	2	1	3	7	∞
B	0	2	1	3	7	∞
D	0	2	1	3	6	∞
E	0	2	1	3	6	7
F	0	2	1	3	6	7

Table 8: Table of the distances for Dijkstra's algorithm

11.2 Programming tips

Due to the fact that the programming of this tool could be sometimes laborious, a couple of tips are provided below for situations where large amount of spatial data are being analyzed.

1) Always try to treat the data sequentially and delete all the variables which are no longer needed. Overloading python's memory slows down the program (or make it crash!).
2) Try minimizing the number of accesses to the database (this part is usually a tradeoff with the first recommendation).
3) When accessing a table with more than 5'000 rows, create indexes on the conditional tuples ("WHERE" clause). Use GiST indexes for the geometries and b-trees for the rest (faster to create than hash tables and more stable).
4) After the creation of new indexes, always use the function "VACUUM ANALZE" of PostgreSQL to make sure the indexes will be used.
5) When selecting a subset of geometric objects, always try to use a box rather than a buffer (so that the geometrical indexes can be used), even if it means selecting more objects than required (it is usually faster to sort them out afterwards).
6) When storing many values in a table, try to sort the values in the tuple which will be accessed later.
7) Prefer integers to strings if attributes in a table need to be accessed.

11.3 Zoom on the visibility analysis

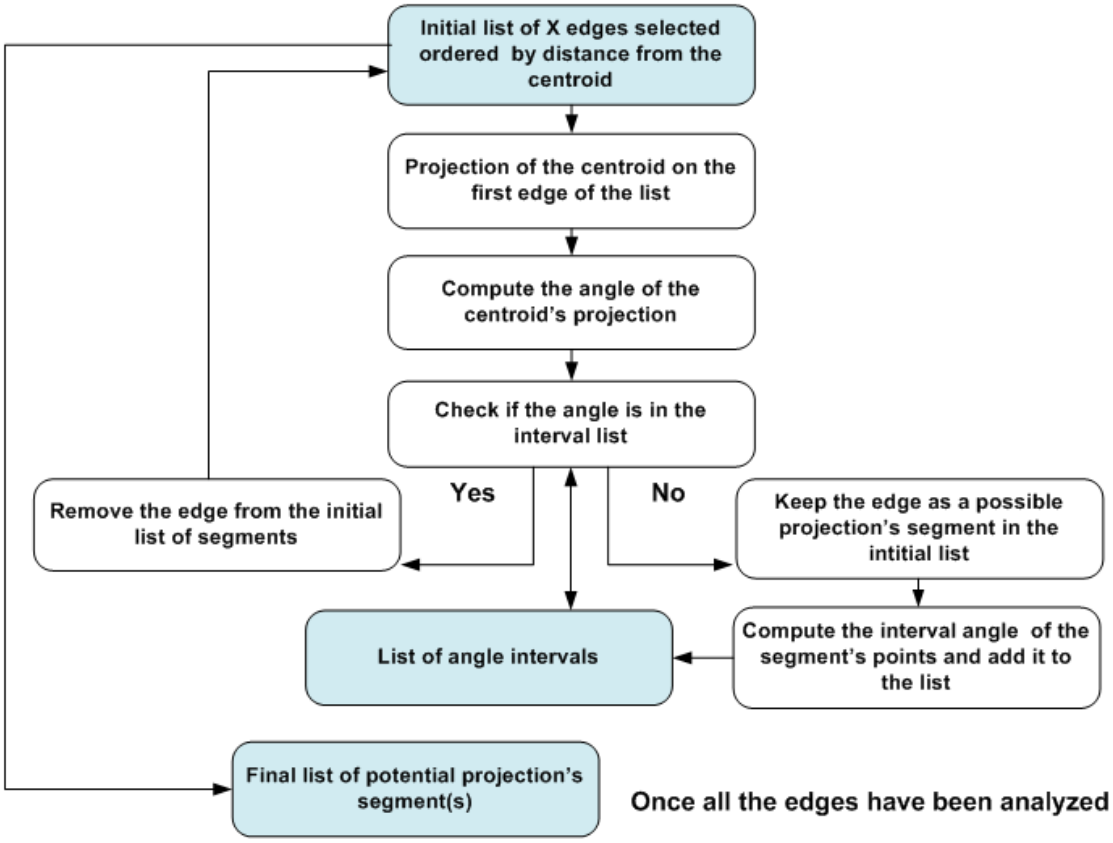

Figure 47: Zoom on the visibility analysis

11.4 Zoom on the distance computation

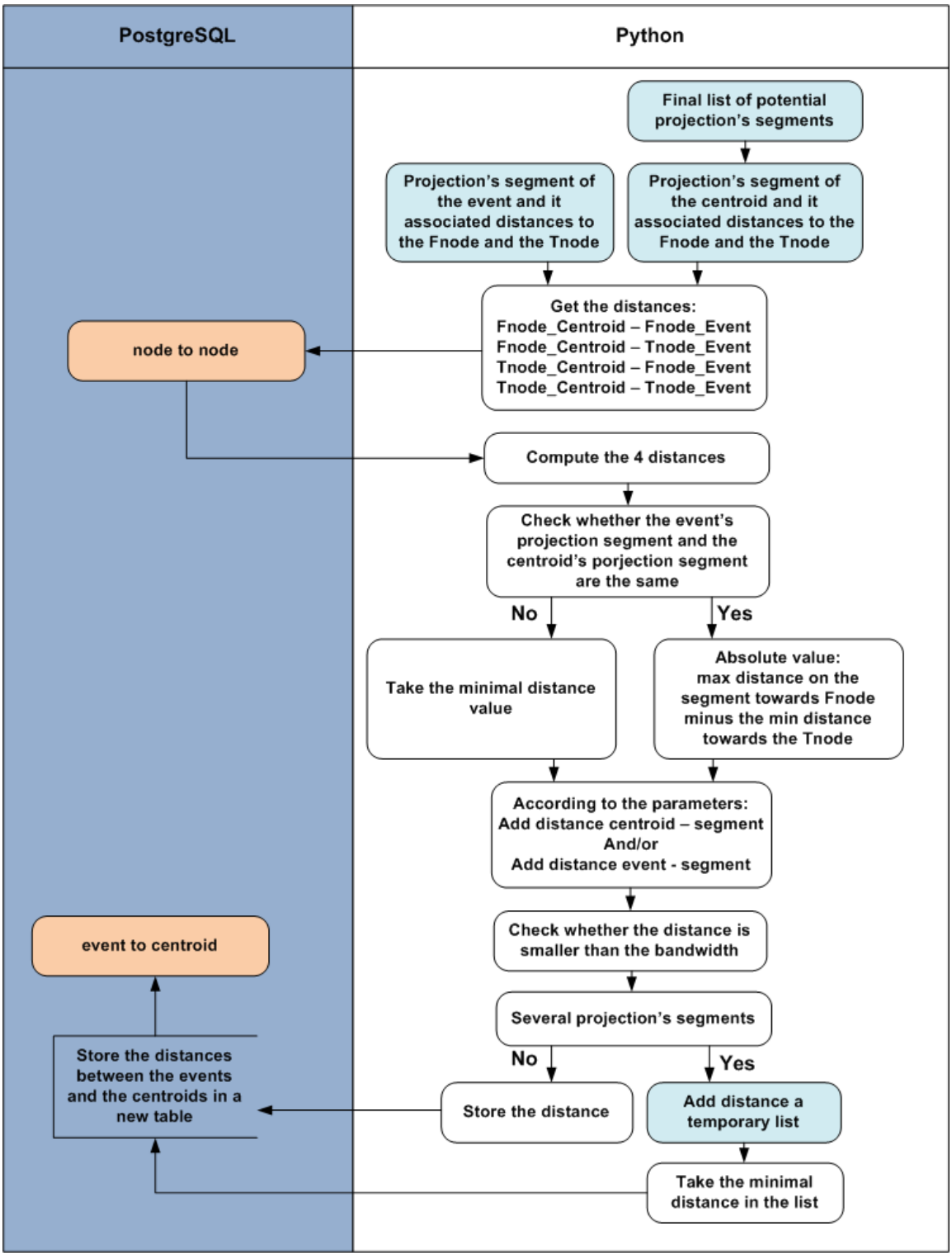

Figure 48: Zoom on the distance computation

11.5 Convex polygons computation

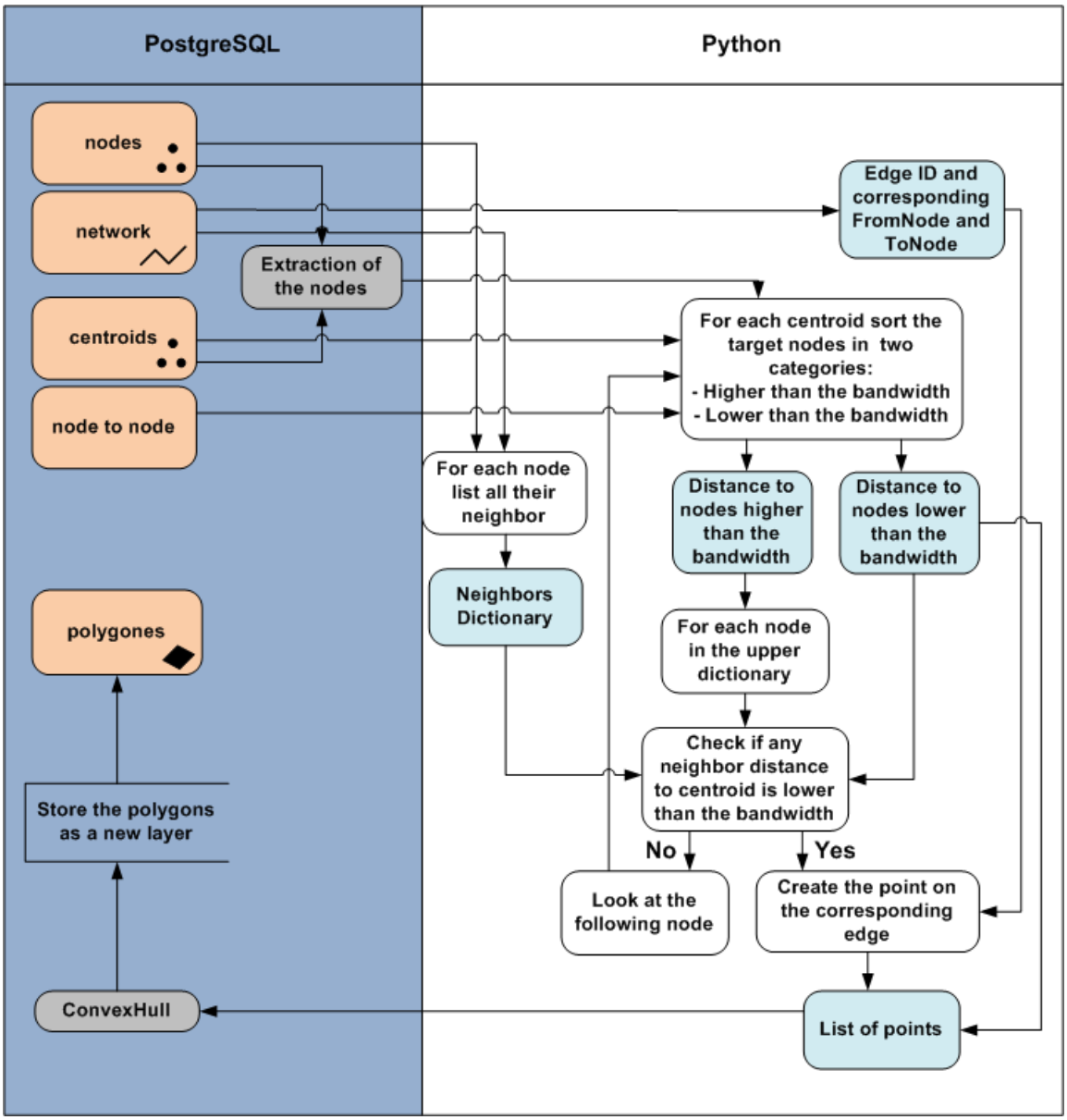

Figure 49: Convex polygons computation

11.6 NetKDE closest on the class retail stores in Geneva

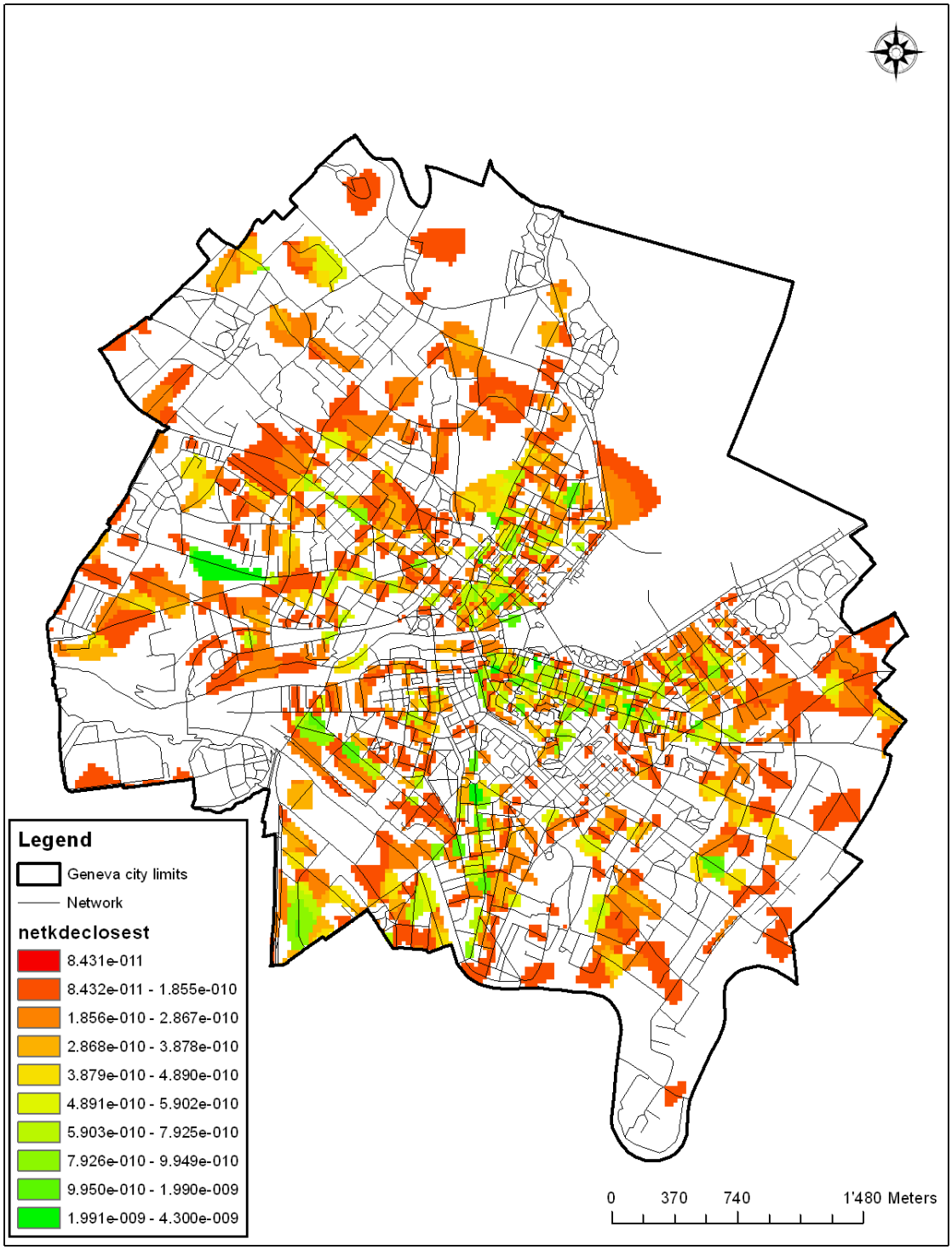

Figure 50: NetKDE closest segment approach on retail stores in Geneva (500 m)

11.7 NetKDE visible on the class retail stores in Geneva

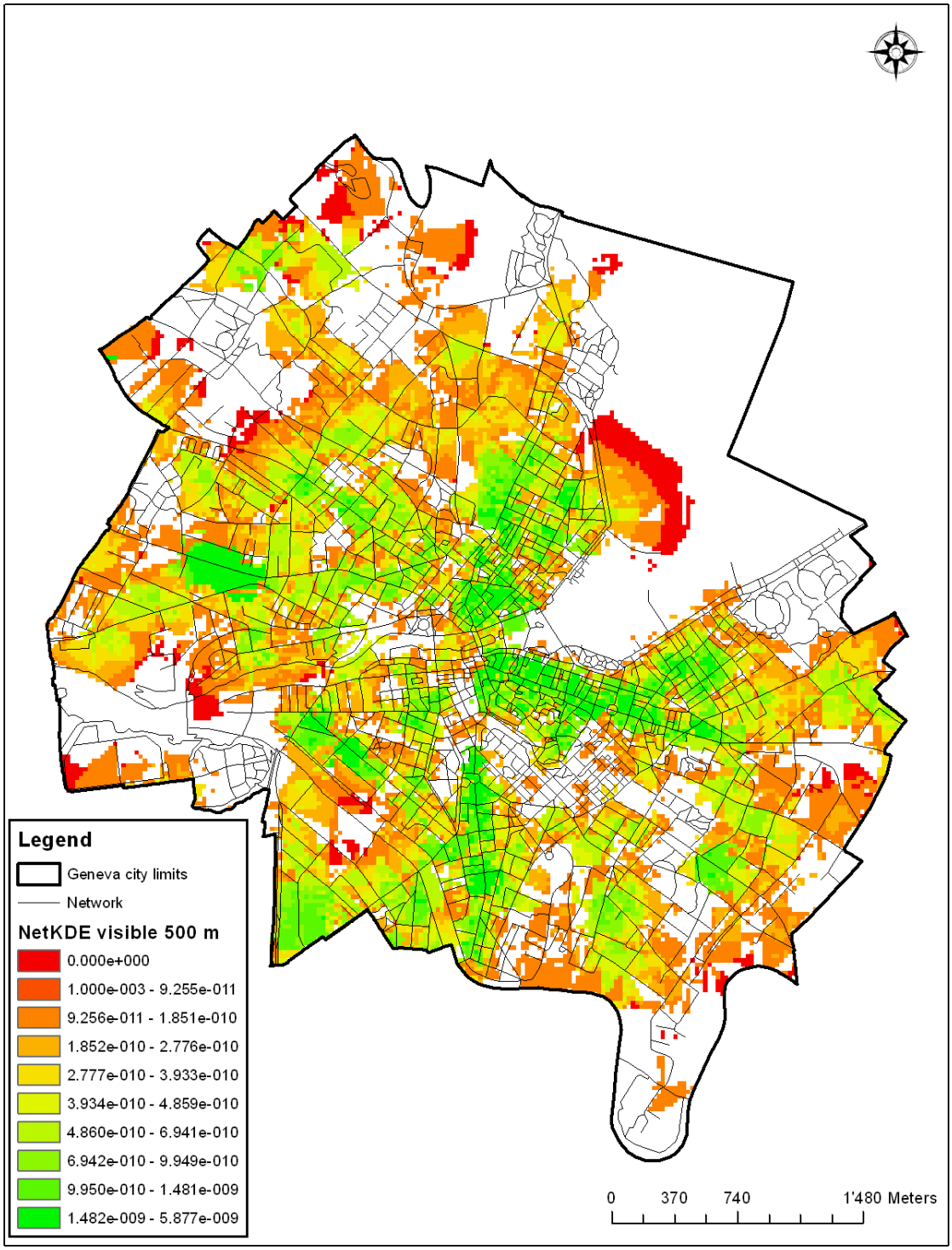

Figure 51: NetKDE visible on the class retail stores in Geneva (500 m)

11.8 KDE on the class retail stores in Geneva

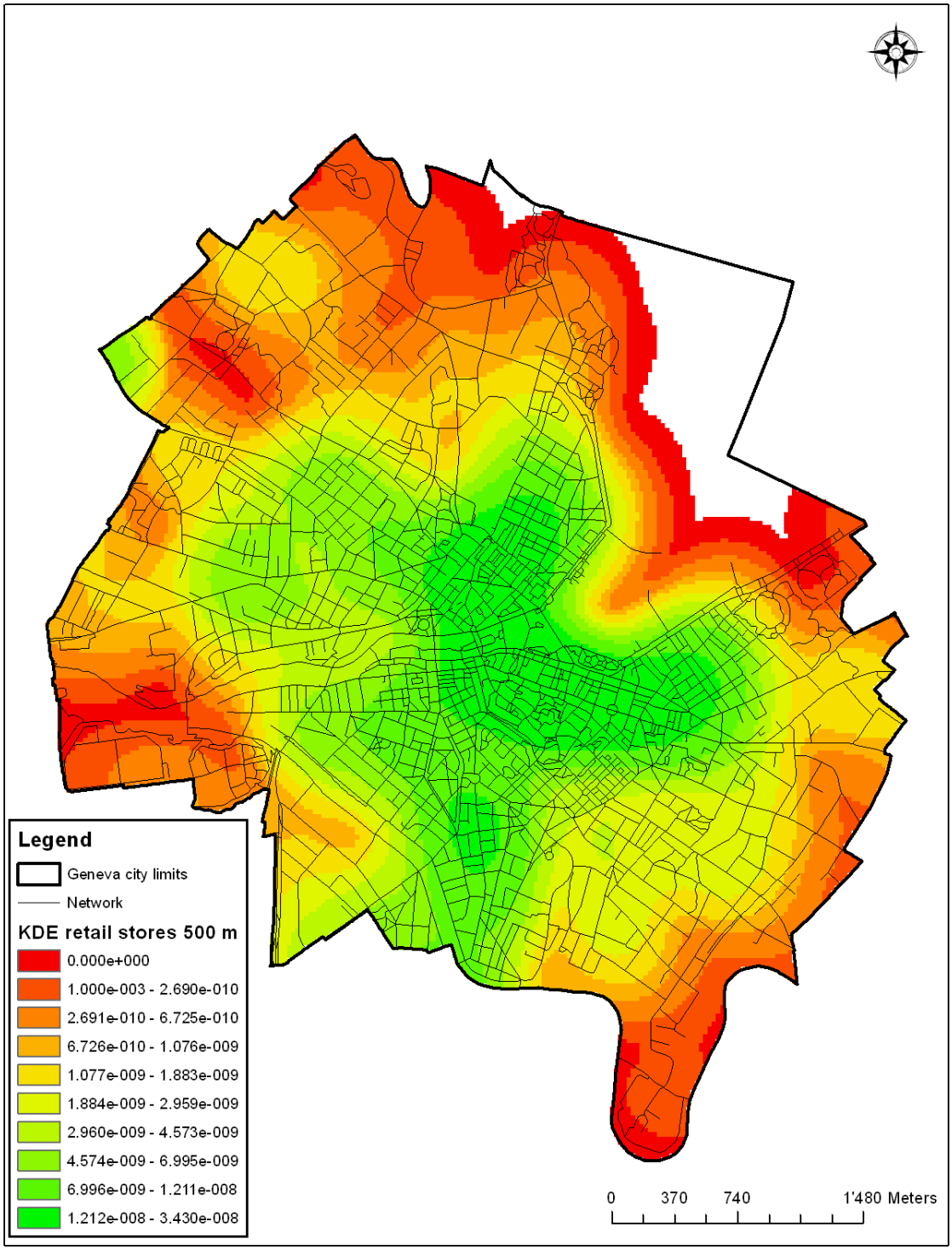

Figure 52: KDE on the class retail stores in Geneva (500 m)

11.9 Density values of economical activities in Geneva

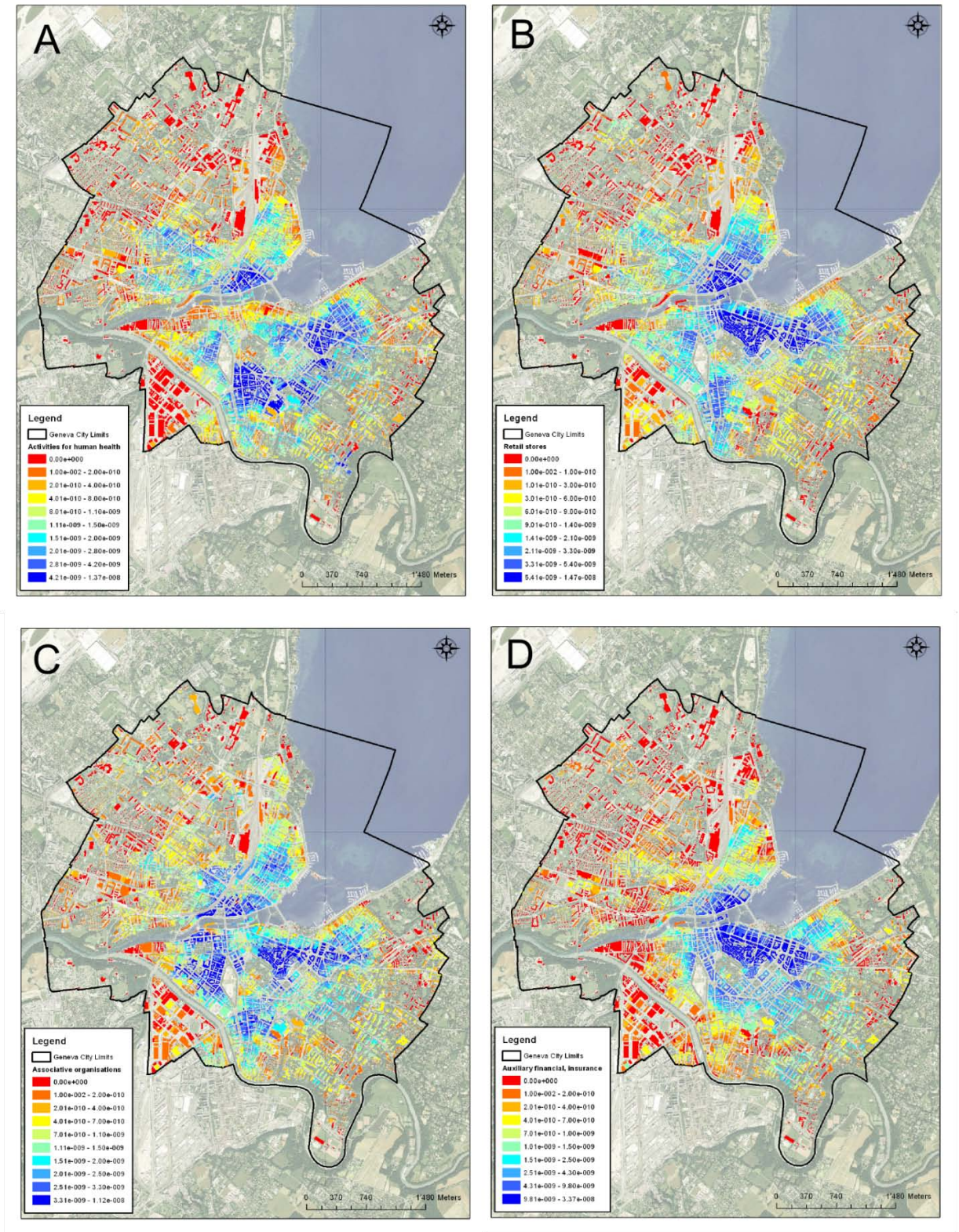

Figure 53: NetKDE 500 meters A) Activities for human health, B) Retail Stores, C) Associative organizations, D) Auxiliary financial and insurance activities

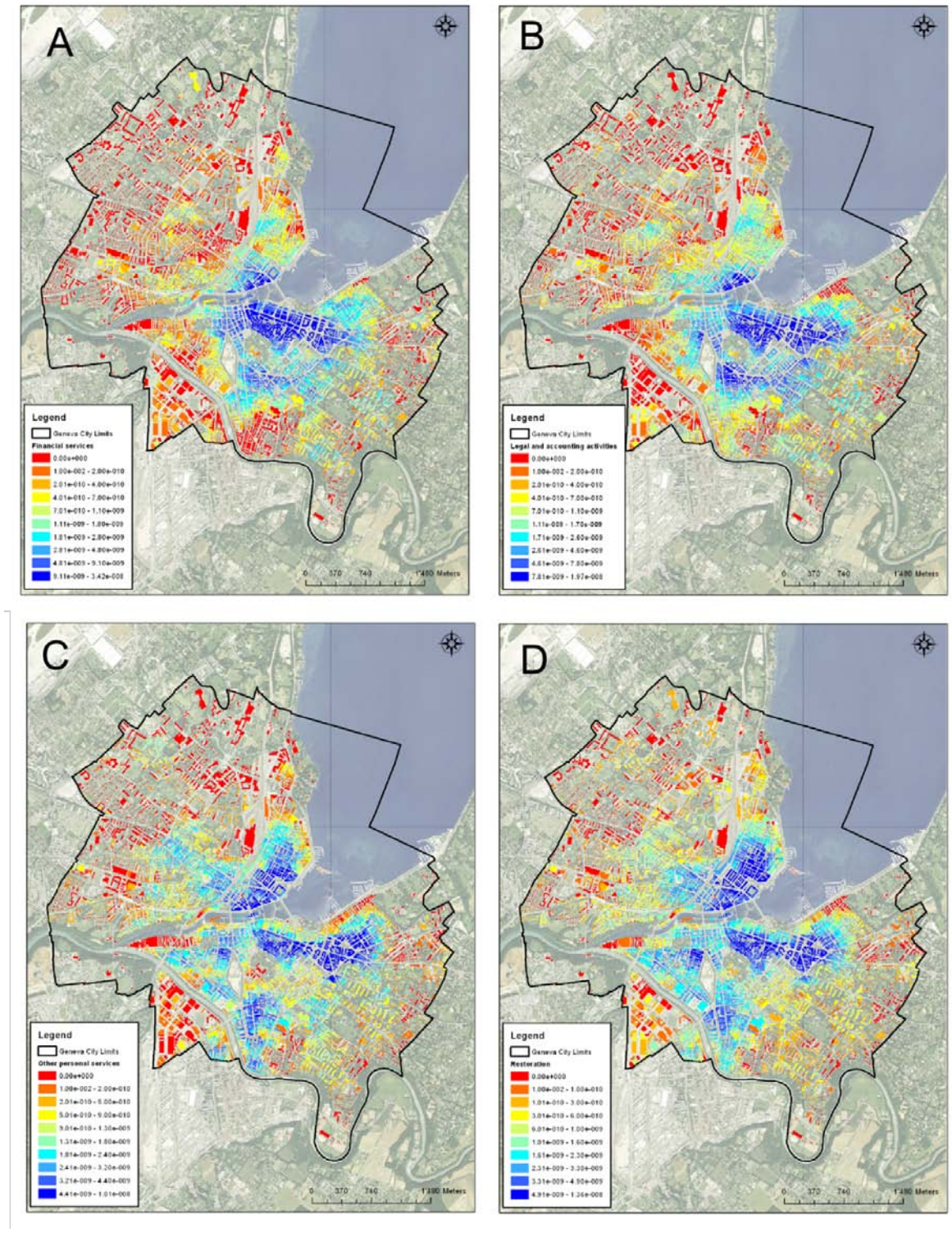

Figure 54: NetKDE 500 meters A) Financial services, B) Legal and accounting activities, C) Other personal services, D) Restoration

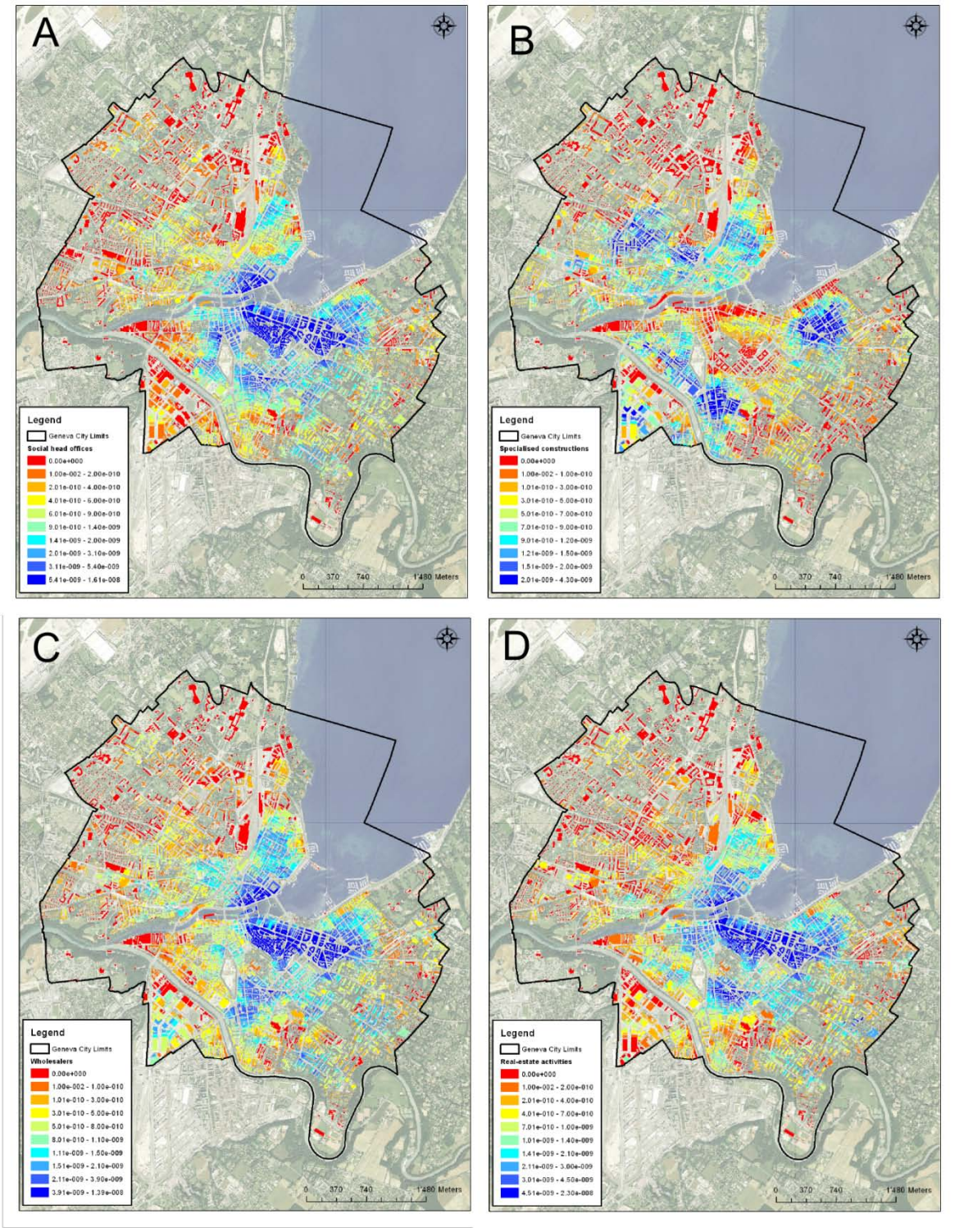

Figure 55: NetKDE 500 meters A) Social head offices, B) Specialized constructions activities, C) Wholesalers, D) Real-estate activities

12 Acknowledgements

Voilà!! J'vais enfin passer au français ! Je voulais d'abord effectuer un remerciement général à tout le personnel du LASIG. J'ai vraiment passé un formidable moment en votre compagnie. Timothée, que te dire sinon que tu m'as été d'une aide précieuse tout au long du projet, et que j'ai grandement apprécié ta compagnie (plus que ta musique). Avec ce projet de Master tu m'as vraiment laissé l'opportunité d'explorer de nouveaux horizons et de faire place à ma créativité (ainsi qu'à la tienne ;), esclave !). Nicolas, je voulais te remercier pour tes conseils avisés ainsi que pour l'intérêt que tu portes à mon travail. Tu m'as vraiment beaucoup aidé à structurer mon projet et mes idées. Je te souhaite plein de réussite pour la fin de doctorat ! Stéphane, ben j'vais arrêter d'écrire parce que tu dois être entrain de t'endormir ! Tu dois me détester pour cette tartine ! Merci pour ta disponibilité à écouter mes idées parfois un peu saugrenues, ça m'a vraiment permis d'avancer. Claudio, merci de t'être proposer de relire ma thèse, même si a priori ce n'est pas forcément ton domaine. Enfin, je tiens à remercier François pour sa compréhension dans les moments difficiles que j'ai dû traverser pendant ce semestre, ainsi que pour ce vol magnifique !